치의과학박사 학위논문

의생명 분야 문서의 언어적/구조적 특징을 이용한 자동 어노테이션에 대한 연구

The Study on Automatic Annotation using

Structural/Linguistic Characteristics of biomedical documents

2015 년 8 월

서울대학교 대학원

치의과학과 의료경영과정보학

남 세 진

의생명 분야 문서의 언어적/구조적 특징을 이용한 자동 어노테이션에 대한 연구

지도교수 김 홍 기

이 논문을 치의과학박사 학위논문으로 제출함
2015년 5월

서울대학교 대학원
치의과학과 의료경영과정보학
남 세 진

남세진의 치의과학박사 학위논문을 인준함
2015 년 7 월

위 원 장 김 명 기 (인)

부위원장 김 홍 기 (인)

위 원 이 재 일 (인)

위 원 이 수 경 (인)

위 원 임 동 혁 (인)

초 록

자동 어노테이션에 대한 연구는 급속도로 증가하는 의생명 분야의 논문과 임상 문서들을 더욱 정확하게 검색하거나 필요한 정보만을 추출할 수 있게 하는 기반이 된다는 점에서 중요하다. 본 연구에서는, 그 중 연구 활동에서 필수적인 논문 검색과 환자의 질병에 대한 진단, 검사, 그리고 처방 등을 기록하는데 필수적인 임상서식의 작성에 초점을 맞추어, 이에 필요한 어노테이션 기술을 연구하였다. 이 두 가지 활동은 의생명 분야의 대표 문서인 논문과 임상서식을 대상으로 일상적으로 일어나는 것이며, 이러한 활동이 효율적으로 개선되는 것은 의생명 분야에서 중요한 의미를 가진다.

먼저, 텍스트 형식의 연구 논문에 대해서는 연구 활동의 방향 설정에 중요한 역할을 하는 초록을 대상으로, 의생명 분야에서 주로 사용하는 IMRAD(Introduction, Methods, Results, and Discussion)로의 자동 태깅을 연구하였다. 이 연구에서는, 기존 언어학 분야에서 의생명 분야의 논문을 대상으로 이룬 결과와 컴퓨터 과학 분야에서 진행돼온 결과를 기반으로, 계산 비용이 적으면서도 높은 성능을 내는 새로운 자동 태깅 시스템을 제안하고 개발하였다. 본 연구에서 제안한 방법을 사용하는 경우, 문장에서 뽑아낸 17개의 특징만으로도 비구조화된 초록을 Accuracy 77.0 ~ 90.3%의 성능으로 분류할 수 있었다. 또한, 기존 연구들에서 사용한 특징들과 함께 사용했을 때는 최대 Accuracy 91.7%의 성능을 보여주었다.

임상 문서의 경우, EMR(Electronic Medical Record)을 시스템을 사

용하는 환경에서는 임상 서식을 통해 생성되는 경우가 대부분이므로, 임상 서식을 대상으로 자동 태깅을 시도하였다. 임상 서식은 연구 초록과는 달리 이미 구조화된 형식을 가지고 있으므로, 본 연구에서는 이 구조 안에 내재된 전문가의 지식을 태깅하고자 하였다. 이를 위해 새로운 지식모델과 이를 이용한 임상 서식 작성 지원 시스템인 STEP(Smart Clinical Document Template Editing and Production System)을 개발하였다. STEP의 시스템의 활용성을 검증하기 위해서는 임상 서식 작성 도구를 개발하여, 지식 모델을 통해 구축된 지식베이스가 임상 서식의 작성을 개선시킬 수 있음을 보였다.

연구 결과는 의생명 분야의 연구자들에게 대규모의 의생명 관련 논문과 임상에서 지속적으로 생산되는 임상 문서가 더욱 정확하게 검색되고 재사용될 수 있음을 보여주고 있다. 이러한 결과는 의생명 분야 전반에서 연구자들의 활동을 개선시킬 수 있다는 점에서 중요하다. 마지막으로, 본 연구의 성과가 다른 연구자들에게도 활용될 수 있도록, 연구 과정에서 추출한 언어 자원과 결과를 확인할 수 있는 시스템을 웹으로 공개하였다.

주요어 : 어노테이션, 구조화된 초록, 임상서식, 문장 분류, 온톨로지

학 번 : 2010-30648

목 차

테이블 목차

그림 목차

부록 목차

I. 서론

1. 연구 배경

의생명 분야의 발전에 따라 병원의 임상 문서와 연구 논문의 수는 급속도로 증가하고 있다. Ware and Mabe (2015) 에 따르면, 2014년 12월 기준으로 심사체계를 갖춘 저널 (28,100개)에서 약 250만 편의 논문이 한 해 동안 출판되었으며, 이 중 의생명 분야 논문들이 전체의 30%를 차지하였다. 의생명 분야 논문의 증가는 NLM(National Library of Medicine)의 논문 인용 데이터베이스 MEDLINE에서도 확인할 수 있는데, (Druss and Marcus 2005)의 분석에 따르면 1994년에서 2001년까지 출판된 논문은 1978년부터 1985년까지 나온 논문에 비해 46%나 증가한 수치를 보여주고 있다. 이와 더불어 실제 환자를 접하는 병원에서도 EMR(Electronic Medical Record)과 EHR(Electronic Health Record)의 도입으로 다양한 임상 문서가 급속도로 증가하고 있다. 이러한 디지털 문서의 증가는 환자에게 양질의 의료 서비스를 제공하기 위해, 진료과정에서 발생하는 진단, 처방, 검사, 비용 처리 등에 관련된 정보들을 디지털화된 형태로 기록하기 때문이다. 미국의 경우, 정부 기관인 CMS(Center for Medicare & Medicaid Services)가 주도하여 EHR 인센티브 프로그램인 “Meaningful Use”를 2010년에 발표하면서(Blumenthal and Tavenner 2010), 병원 정보의 디지털화와 이를 통한 서비스 향상을 추진하고 있다. 국내의 경우는 2002년도 의료법 제 23조의 개정으로 전자문서 형태의 기록이 가능하게 되었으며, 이 후 서울대학교병원을 포함한 대형병원과 1,2차병원의 대부분이 EMR시스템을 도입하여 사용하고 있어

(이경진 2010) 디지털화된 임상 문서의 증가는 세계적인 추세가 되어가고 있다. 따라서, 의생명 분야의 문서 증가 현상은 연구 분야에서부터 환자를 다루는 병원에 이르기까지 보편적인 현상으로 자리잡고 있으며, 연구자와 의료진들이 접할 수 있는 정보의 양 역시 증가하고 있음을 유추할 수 있다.

그런데, 의생명 분야 정보의 양적 증가는 정보의 분열(Fragmentation)을 야기한다는 점에서 중요한 문제점으로 인식되어 왔다. 분열 현상은 분야간 논문의 인용이 줄어들거나 없어지는 현상으로 나타나는데, 이러한 현상은 연구 분야 간이나 특정 연구 분야를 구성하는 세부 분야, 또는 연구 대상 및 방법에서도 나타난다(Ganiz, Pottenger et al. 2005). Swanson은 분열 현상이 의생명 분야의 문서 규모가 사람의 처리 능력을 벗어남에 따라, 연구자들이 어쩔 수 없이 자신과 관련된 연구 분야에만 대응하려는 성향을 가지기 때문임을 지적하였다(Swanson 2001, Bruza and Weeber 2008). 의생명 연구가 최종적으로는 질병의 예방, 진단, 치료를 위한 것이라고 했을 때, 연구자들이나 의료진들이 대규모의 문헌에서 목적에 부합하는 정보를 빠르고 정확하게 발견하는 것과 기존의 정보를 효율적으로 재사용할 수 있게 해주는 것은 더욱 더 중요해지고 있다. 텍스트 마이닝(Text Mining)은 바로 이러한 필요를 해결하기 위한 연구 분야로, 컴퓨터를 이용하여 대규모 텍스트에서 새로운 지식이나 내재된 지식을 밝혀내는 것을 목적으로 한다(Hearst 1999). 일반적으로 텍스트 마이닝은 대규모 문헌들에서 관련된 문서만을 얻기 위한 정보 검색 단계, 검색된 문서에서 의미 있는 정보를 자동으로 태깅하는 어노테이션 단계,

그리고 이러한 추출된 정보들에서 관계를 분석하여 새로운 지식을 발견하는 단계로 구분될 수 있으며, 각각의 단계에서 다양한 연구 성과와 응용 프로그램들이 꾸준히 발표되고 있다 (Cohen and Hersh 2005, Ananiadou and McNaught 2006, Zweigenbaum, Demner-Fushman et al. 2007, Fleuren and Alkema 2015).

어노테이션 단계에서 부착되는 태그는 문서 검색과 정보 추출에서 중요한 색인정보로 사용되어, 의생명 분야에서는 그동안 다양한 수준의 자동 어노테이션 기술이 연구되어 왔다. 문서에 부착되는 태그는 선택된 텍스트의 문법적, 의미적, 또는 구조적인 정보를 명시적으로 표현하여, 기존의 문서를 색인하거나 보관할 때 더욱 더 많은 정보를 사용할 수 있기 때문이다. 이중 가장 일반적으로 사용되는 것이 문서에 포함된 단어에 개념명을 부착하는 개체명 인식(Named Entity Recognition)이다. 개체명 인식은 문서에 있는 유전자, 단백질, 또는 질병 등을 표현하는 단어에 'Gene', 'Protein', 또는 'Disease'와 같은 태그를 부착하여, 사용자가 높은 정확률로 문서를 검색하게 하거나 유전자나 질병과 같은 개체간의 관계를 추출하기 위한 전처리 역할을 한다. 이와 같은 어노테이션의의 유용성으로 인해, 최근에는 단어를 대상으로 한 어노테이션에서 벗어나, 특정한 의미를 가지는 문장이나 연속된 문장 집합, 또는 구조를 대상으로 하는 자동 어노테이션 연구가 진행되고 있다(Wilbur, Rzhetsky et al. 2006, Groza, Hassanzadeh et al. 2013).

자동 어노테이션에 대한 연구는 급속도로 증가하는 의생명 분야의 논문과 임상 문서들을 더욱 정확하게 검색하거나 필요한 정보만을 추출할 수 있게 하는 기반이 된다는 점에서 중요하다. 본 연구에서는, 그 중 연구 활

동에서 필수적인 논문 검색과 환자의 질병에 대한 진단, 검사, 그리고 처방 등을 기록하는데 필수적인 임상서식의 작성에 초점을 맞추어, 이에 필요한 어노테이션 기술을 연구하였다. 이 두 가지 활동은 의생명 분야의 대표 문서인 논문과 임상서식을 대상으로 일상적으로 일어나는 것이며, 이러한 활동이 효율적으로 개선되는 것은 의생명 분야에서 중요한 의미를 가진다.

2. 연구 목적

본 연구는 연구자나 의료진들의 연구활동 및 임상서식 작성 과정에서, 효과적으로 논문이나 서식을 검색할 수 있도록 검색에 필요한 태그를 문서에 자동적으로 부착하는 것을 목적으로 한다. 의생명 분야 문서의 효과적인 검색이라는 목표를 위해, 연구는 세부적으로 두 가지 방향에서 이루어졌다. 첫 번째, 연구활동에서 필수적인 논문검색이 효율적으로 이루어지도록 의생명 분야의 비구조화된 초록에 IMRAD(Introduction, Methods, Results, and Discussion) 의 섹션명을 자동적으로 태깅하는 연구를 수행하였다. 이를 통해 연구자는 관련된 논문을 검색하고 자신의 목적에 부합하는 논문임을 결정하는 과정에서, 연구 목표나 방법, 또는 결론과 같은 특정 섹션만 확인할 수 있어 빠른 판단이 가능하다. 두 번째, 의료기관에서 임상서식을 작성하는 의사, 간호사, 또는 의무기록사에게 기존 임상서식에 내재된 지식들이 재사용될 수 있도록, 서식의 내용을 온톨로지를 이용하여 태깅하는 연구를 수행하였다. 임상 서식은 전문가에 의해 설계된 문서라는 점에서 문서를 구성하는 필드나 배치에 전문가의 지식이 내재되어 있다고 볼 수 있는데, 본 연구에서는 이러한 내재된 지식이 온톨로지를 이용하여 명시적으로 태깅되어 재사용될 수 있도록 새로운 지식 모델과 온톨로지를 정의하였다.

본 연구는 의생명 분야의 대표적인 두 문서인 연구 논문과 임상서식을 대상으로 자동 어노테이션 연구를 수행하였다는 점에서 의미 있는 시도이다. 이 두 종류의 문서는 구조적인 차이가 있으나 의생명 분야의 대표적인 문서들이다. 본 연구에서는 이들 문서를 태깅할 때에, 문서에 내재된 의생

명 분야 및 사용 분야의 특징을 추출하고 이를 활용하고자 노력하였다.

3. 논문의 구성

본 논문은 어노테이션 대상 문서에 따라, 크게 두 파트로 나누어진다. 첫 번째 파트는 2장, 3장, 4장으로 구성되며, 초록을 대상으로 한 어노테이션 연구를 소개한다. 먼저, 2장에서는 논문 초록을 자동 어노테이션하는데 있어 초록에 나타나는 언어학적 특징을 추출하기 위해, 대규모의 구조화된 초록을 분석하고 각 섹션을 대표하는 특징을 분석하는 과정을 설명한다. 3장에서는 2장에서 추출한 언어학적 특징을 이용하여 비구조화된 초록들을 태깅하기 위한 자동 분류 시스템을 설명한다. 4장에서는, 3장에서 개발한 자동 분류 시스템을 이용한 웹 서비스를 소개한다. 이 검색 서비스를 통해 논문의 초록이 자동으로 태깅되었을 때에 기존의 검색 서비스가 어떻게 개선될 수 있는지를 설명한다.

두 번째 파트는 5장, 6장으로 구성되며 임상서식을 대상으로 한 어노테이션 연구를 소개한다. 5장에서는 임상서식에 내재된 전문가 지식을 모델링 하기 위해 제안한 지식 모델과 온톨로지를 소개하고, 이를 통해 임상서식의 필드들이 의미적으로 연결되어 저장, 관리될 수 있음을 설명한다. 6장에서는 5장에서 구축한 온톨로지 인스턴스들을 이용한 임상 서식 지원 시스템 STEP(Smart Clinical Document Template Editing and Production System)을 설명한다. 마지막으로 7장에서는 본 연구가 가지는 의미, 기여, 그리고 연구의 한계에 대해서 설명한다.

Ⅱ. 구조화된 초록의 언어적 특징 추출

1. 연구 배경

초록은 연구자가 논문의 내용을 파악할 수 있도록 그 내용을 간략하게 요약한 문장들이다(사공철, 김종천 et al. 1996). ISO/TC 46[1]에서는 초록을 지시적 초록(Indicative Abstract), 정보적 초록(Informative Abstract), 그리고 이 둘을 혼용한 정보-지시적 초록(Informative-indicative)으로 구분한다. 지시적 초록은 논문에서 무엇을 설명할 것인지를 기술하는 방식이고, 정보적 초록은 논문 전체의 내용을 요약하여 설명하는 방식으로 주로 연구 목적, 방법, 결과를 포함한다. 마지막으로 이 둘을 혼용한 방법은 중요한 부분에서는 정보적 방식을 사용하고 나머지 부분은 지시적 방식을 따르는 경우이다. (van_der_Tol 2001, Budgen, Burn et al. 2011)은 초록의 역할을 4가지로 설명하고 있는데, 첫 번째, 초록은 독자로 하여금 논문을 더 살펴볼지에 대한 판단을 돕는다. 두 번째, 초록은 논문 전체의 내용을 요약하여 독자에게 필요한 정보를 제공한다. 세 번째, 독자가 논문 전체 내용이나 특정 부분을 읽는데 도움을 줄 수 있는 논리 구조를 제공한다. 네 번째, 요약과 함께 제공하는 키워드를 통해 논문 색인에 필요한 정보를 제공한다.

그러나, 초록이 위와 같은 역할을 한다 해도 표준화된 형식의 부재로 논문의 내용을 이해하기에는 부족한 정보를 제공할 수 있는데, 이러한 경우

[1] http://www.iso.org/iso/catalogue_detail.htm?csnumber=4084

독자는 논문이 자신이 찾는 논문임을 판단하기 위해서 본문 중에서 관련된 부분을 찾아서 읽어야 하는 번거로움이 생긴다(Budgen, Burn et al. 2011). 이러한 문제에 대해서, 의생명 분야에서는 초록에서 제공하는 정보의 가치와 독자의 가독성을 개선시키는 방안으로 구조화된 초록을 사용하고 있다. 구조화된 초록은 연구 내용을 명확하게 구분하여 작성할 수 있는 형식을 제공하여, 연구자들이 논문의 내용을 조직적으로 요약할 수 있게 도움을 뿐 아니라 관련된 논문을 보다 쉽게 검색, 선택, 또는 추출할 수 있게 할 수 있는 장점을 가진다(Harbourt, Knecht et al. 1995, Bayley and Eldredge 2003, Gerstein, Seringhaus et al. 2007, Budgen, Burn et al. 2011, Ripple, Mork et al. 2011, Zhang and Liu 2011).

본 연구는 이러한 구조화된 초록의 장점을 비구조화된 초록에도 적용될 수 있도록, 비구조화된 초록을 구조화된 것으로 자동적으로 태깅하는 것을 목적으로 한다. 이러한 목적을 달성하기 위해서는 일반적으로 분류시스템을 사용하는데, 분류 시스템의 성능은 크게 분류 알고리즘의 성능과 어떤 특징들을 이용하였는지에 영향을 받는다. 본 연구에서는 분류 시스템에 사용할 특징을 선택함에 있어, 의생명 분야의 특성을 잘 나타낼 수 있는 것을 찾아 성능을 개선시키는 방향을 선택하였고, 특히 의생명 분야의 구조화된 초록에서 나타나는 언어적 특징에 초점을 맞추었다. 이를 위해 본 장에서는 의생명 분야의 구조화된 초록에서 나타나는 언어적 특징을 분석하고, 이를 분류 시스템에서 사용할 수 있는 형태로 변환하는 내용을 설명한다.

2. 연구 목적

본 연구에서는 구조화된 초록의 형식 중에 가장 많이 사용되는 IMRAD 형식(Huth 1987, Sollaci and Pereira 2004)을 기반으로, 각 섹션의 특징을 잘 나타내는 언어적 특징을 추출하고 분석하였다. 추출된 특징들은 비구조화된 초록을 IMRAD의 섹션명으로 태깅하는 시스템에서 사용되기 위해, 특정 섹션에서의 중요도를 계산한 후 가중치 테이블에 기록하였다.

3. 관련 연구

논문을 대상으로 한 언어학 분야의 연구는 논문의 문장들이 저자가 전달하려는 메시지에 영향을 받는 언어적 특징이 있음을 보여주고 있다. 그 대표적인 특징이 동사의 종류나 형식, 그리고 연속된 단어로 구성된 N-gram 이다.

그 중 동사는, 동사의 종류, 태, 시제, 그리고 조동사나 To부정사의 사용여부가 문장을 구별하는데 중요하게 사용되고 있다. (Hanania and Akhtar 1985)는 자연 과학 분야의 석사학위 논문 20편을 5개의 섹션, “Introduction”, “Review”, “Methods”, “Results”, 그리고 “Discussion”으로 구분하고, 각 섹션에서 사용된 동사의 형태를 분석하였다. 이 연구에서 제시한 중요한 발견은 섹션별로 동사의 사용에 중요한 차이점이 있다는 것이다. 예를 들어, “Methods” 섹션에서는 동사가 수동태와 과거형으로 주로 사용되었다는 것, 현재형의 동사는 “Introduction”에서 빈번하게 사용되었다는 것, 그리고 조동사는 “Methods” 섹션에서는 가장 적게 사용되었으나 “Discussion”섹션에서는 가장 많이 사용되었

다는 것 등이다. 또한, (Williams 1996)은 동사들을 “reporting”, “observation”, “relations”, “defining”, “cause and growth”, 그리고 “methods”등의 그룹으로 분류하고, 이 부류의 동사들이 IMRAD 섹션에서 어떻게 사용되는지를 분석하여 흥미로운 결과를 보여주었다. “Observation” 에 포함되는 동사 “show”, “present”, 그리고 “follow”는 주로 능동태로 사용되었고, 같은 부류의 동사 “find”, “observe”, “see”, 그리고 “demonstrate”는 거의 수동태로 사용되었다. 또한 “Suggest”, “Report”, 그리고 “Confirm”등과 같은 “Relations” 부류의 70%에 해당하는 동사들이 “Results”와 “Discussion”섹션에서 사용되었음을 발견하였다.

(de Waard and Pander Maat 2012)의 연구에서는 문장에 포함된 동사의 시제나 법(mood), 연구 문제를 정의할 때 사용되는 부정적인 표현, 또는 연구의 의미를 설명하기 위한 표현들의 형식을 변경하였을 때, 사용자가 문장의 의미를 어떻게 해석하는지를 실험하였다. 이들의 연구 결과 중 본 연구에서 관심을 가진 것은 동사의 시제와 조동사의 사용이 문장의 내용을 해석하는데 영향을 준다는 점이다.

N-gram 역시 문장 분류에서 중요하게 사용된다. N-gram은 “lexical bundle”, “multi-word patterns”, “formulaic patterns”, “clusters”, 또는 “indicator phrase”, “collocation”으로도 불리며, 말뭉치에서 빈번하게 사용되는 단어들의 조합으로 정의되어왔다(Biber, Conrad et al. 2004). 비록 N-gram은 문장에서 문법적으로 명확하게 구분되는 것은 아니지만, 문서의 내용을 표현하는데 가장 기본적인 구성요소의 역할을 한다. 몇몇 연구

들에서는 이러한 N-gram이 연구 논문에서 어떤 특징을 가지는지 보고하고 있다. (Hyland 2008)는 대규모의 연구 논문, 박사 및 석사 학위 논문에서 N-gram의 사용 행태를 분석하여 N-gram을 의미에 따라 “Research-oriented”, “Text oriented”, 그리고 “Participant-oriented”로 구분하였다. “Research-oriented”는 실세계의 경험이나 활동을 구조화할 때 사용되는 표현으로 Location (예, at the beginning of, at the same time, in the present study), Procedure (예, the use of the, the role of the, the purpose of the, the operation of the), Quantification(예, the magnitude of the, a wide range of, one of the most), Description (the structure of the, the size of the), 그리고 Topic (in the Hong Kong, the currency board system)등과 같은 세부 부류를 포함한다. “Text-oriented” 부류는 전달하려는 메시지에 맞게 텍스트를 구성할 때 사용하는 표현으로 Transition signals (on the other hand, in addition to the, in contrast to the), Resultative signals (as a result of, it was found that, these results suggest that), Structuring signals (in the present study, in the next section, as shown in fig.), framing signals (in the case of, with respect to the, on the basis of, in the presence of, with the exception of)등의 세부 부류를 포함한다. 마지막으로 “Participant-oriented” 부류는 텍스트의 저자나 독자를 위한 것으로 stance features (are likely to be, may be due to, it is possible that)와 engagement features (it should be noted that, as can be seen)등의 세부 부류를 포함한다(Lorenzo Salazar 2011). (Csomay 2012)의 연구에서는 구어

체 표현의 코퍼스에서 담론의 구조와 N-gram들간의 관계를 분석하여, 담론에 따라 각기 다른 n-gram들이 사용되고 있음을 발견하였다. 즉, 담론의 특성을 나타내는 n-gram들이 존재하고 있다는 것이다. (Cortes 2013)의 경우는, 연구 논문의 “Introduction”섹션에서 저자가 전달하려는 메시지와 N-gram간의 관계를 분석한 사례이다. Cortes는 “Introduction” 섹션에서 연구 목적을 설명하기 위해 사용되는 N-gram들을 발견하였는데, “the purpose of this study was”가 그 중의 하나이다.

(Ron Daniel 2012)의 연구는 초록에 있는 연구성과와 관련 있는 내용을 추출하기 위해 N-gram을 사용한 경우이다. N-gram만을 사용해서 F-score 70%의 성능을 보여주어, 기계학습을 이용한 방법보다 적은 계산비용으로 정보 추출이 가능함을 보여주었다. 이러한 연구 결과는 N-gram이 문장을 분류하는데 좋은 단서로서 사용될 수 있고, 나아가 문장 분류의 성능을 개선시킬 수 있음을 보여준다.

4. 연구 방법

본 장에서는 언어학 분야에서 이루어진 이상의 성과를 초록의 문장을 IMRAD로 태깅하려는 연구 목적에 적용될 수 있도록, Figure 1과 같은 절차에 따라 추출한 각 섹션의 언어적 특징을 설명하고, 언어적 특징을 중심으로 분석한 내용을 설명한다.

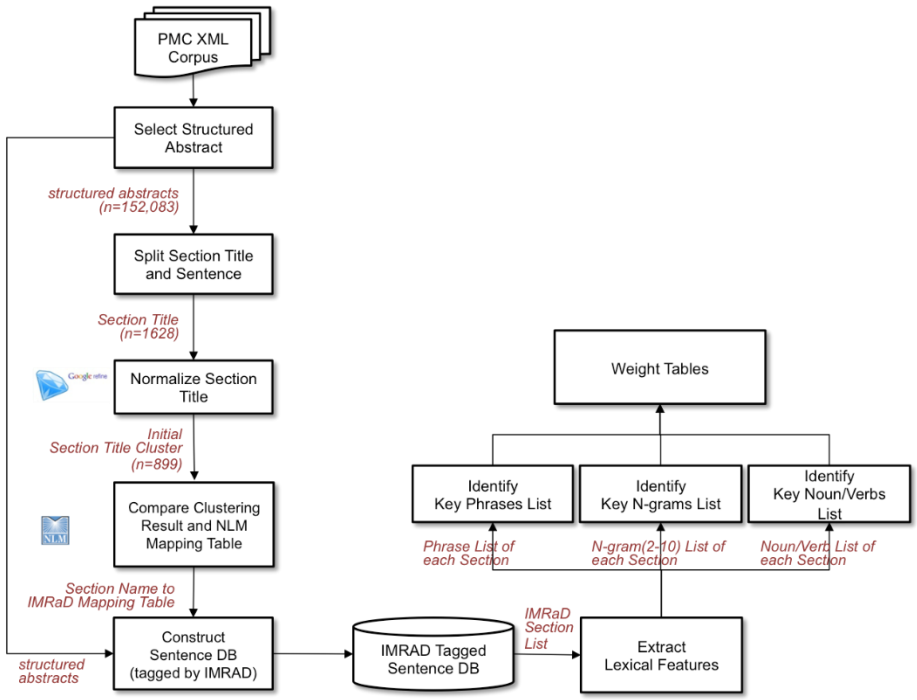

Figure 1 섹션의 언어적 특징 추출 절차

4.1. 데이터 코퍼스

본 연구에서 사용한 코퍼스는 PubMed Central (PMC)에서 데이터 마이닝을 위해 제공하는 Open Access Subset이다. 코퍼스는 총 536,682개의 초록으로 구성되어 있으며, XML (Extensible Markup Language) 형식으로 기록되어 있다. 이들 중 본 연구의 대상이 되는 구조화된 초록은 160,150개으로, 전체 코퍼스의 29.73%에 해당하는 분량이었다. 이들 중 논문의 타입이 “research-type”인 것들만 선택하였고, 그 수는 총 152,083개이었다. Figure 2에서 보는 바와 같이 "research-type"은 전체 구조화된 초록 중 95.1%에 해당하였다. 나머지 비구조화된 초록이면

서 research-type의 초록 221,261개는 분류시스템을 평가하기 위한 테스트 집합을 만들기 위해 사용되었다.

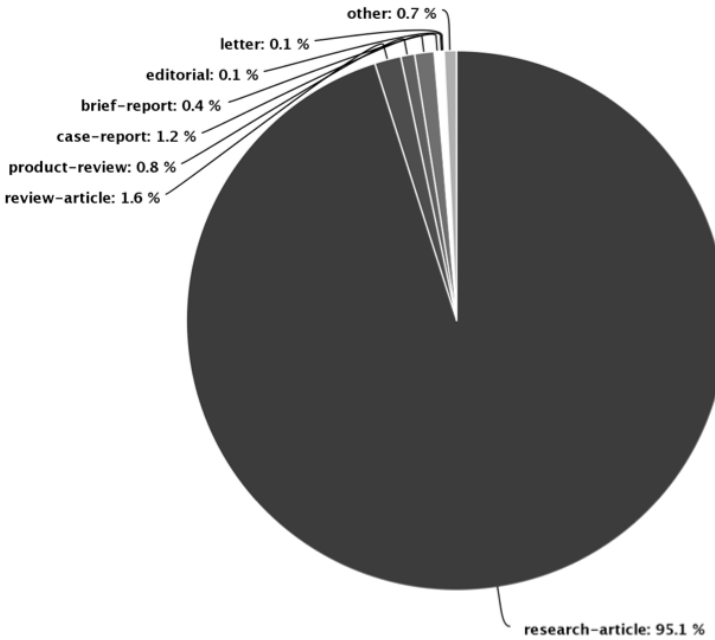

Figure 2 데이터 코퍼스의 Article 타입 분포

4.2. 섹션 정규화

152,083개의 구조화된 초록에서 IMRAD의 각 섹션에 해당하는 문장을 분리하기 위해 XML parser를 이용한 프로그램을 작성하였으며, 이를 이용하여 초록에서 섹션명과 섹션에 포함된 문장들을 추출하였다. 추출된 섹션명은 총 1,628 종류가 있었는데, 이렇게 많은 수의 섹션명이 사용된 이유는 섹션명들이 단 복수 표현의 차이 (“Conclusion”, “Conclusions”), 수식어의 사용 (“Conclusion”, “Major Conclusion”), 사용된 단어의 순서의 차이 (“Conclusions and Significance”,

"Significance and Conclusion"), 그리고 두 섹션명이 같이 사용("Methods and Result", "Result and Discussion") 되었기 때문이었다. 다양한 섹션명은 데이터 정제 및 변환 도구인 Open Refine[2]을 사용하여 그룹화 하였으며, 그 결과 1,001개의 섹션명으로 줄일 수 있었다. Table 1은 그룹화한 결과 중 대표적인 10개를 보여주고 있다.

Table 1 주요 섹션과 변형들

Introduction	Introduction, Introduction and background, Introduction and Design, Introduction and hypothesis, Introduction and methods, INTRODUCTION AND OBJECTIVE, INTRODUCTION AND OBJECTIVES, Introduction/Methods, INTRODUCTION/OBJECTIVES
Background	Background, Background and Objectives, Background and Aims, Background/aims, Background/aim, BACKGROUND AND OBJECTIVE, Background and Purpose, Background and Aim, Backgrounds, Background and Methods, Background/objectives, Background & Aims, Background/Purpose, Background/Objective, "Background, aim, and scope", Background and Methodology, Background and hypothesis, Backgrounds and Aims, Backgrounds and aim, Background/hypothesis, Background/Abstract, Background, Aims of the Study, Background data, Background Context, Background and study objective, Background and significance, Background and rationale, Background and object, Background and motivation, Background and case presentation, Background & Objective, Background & aim
Purpose	Purpose, Purpose and Methods, Purpose and Principal Findings, Purpose of Review, Purpose of study, Purpose of the Study, Purpose/Methods, Purposes

[2] http://openrefine.org/

Aim	Aims, Aim, Aims/hypothesis, Aims and objectives, AIM OF THE STUDY, Aims of the study, Aim and Objectives, Aim/Objective, Aims and Objective, Aims and methods, Aims/Objectives, Aim and Method, Aim and methods, Aim and scope, Aim of study, Aim/Background, Aim/hypothesis, Aim/Methods, Aim/Objectives, Aims and approaches, Aims and background, Aims&methods, Aims/Introduction
Object	Objective, Objectives, Objectives and Methods, Objective and methods, Objectives/methods, Object, Objects, Objective and design, Objective of review, Objective of the study, OBJECTIVES AND BACKGROUND, Object or purpose of study, Objective and background, Objective and Discussion, Objective and study design "Objective, Materials and Methods"
Design	Design, Design and setting, Design and methods, Design and participants, Design/Methods, Design and patients, "Design, Setting, and Participants", Design/Methodology/Approach, Design and measurements, "Design, setting and patients", DESIGN/MEASUREMENTS, Design & Setting, Design and Method, Design and subjects, Design of study, Design process, DESIGN STUDY, "Design, Materials and Methods", "Design, participants, and methods", "Design, participants, measurements", "Design, setting and participants", "Design, setting, participants", "Design, Setting, Participants & Measurements", "Design, settings, material, and measurements", "Design, subjects and measurements", Design/ Methods, Design/Method, Design/Population, Design/Setting/Patient population, Design/subjects
Methods	Methods, Methodology/Principal Findings, Method, Methods and findings, Methods/Design, Methodology, Methods and results, Methodology and Principal Findings, Methods and design, Methods/Principal Findings, Methods and Principal Findings, Methodology/Principle Findings, Methodology/Findings, Methodology and Findings, Methodology/Principal Finding, Methods/Results, Methods/Findings, Method/Design, Methods and materials, Methodology

	and results, Methodology/ Principal Findings, Methodology/Results, Method and results, Methodology and Principle Findings, Methods & Results, Methods and Material, Method and Findings, ...
Result	Results, Results and Discussion, Result, Results and conclusion, Results and conclusions, Results & Discussion, Results & Conclusion, Result and Conclusion, Results/Conclusion, Result(s), Result and Discussion, Results and discussions, Result & conclusion, Results & Conclusions, RESULTS AND DISSCUSSION, Results and Interpretation, Results and Methodology, Results and Observations, Results and Outcome, Results and Principal Findings, Results of comparative analysis, Results/ Discussion, Results/Conclusions, Results/Significance
Discussion	Discussion, Discussion and Conclusion, Discussion and Evaluation, Discussion and conclusions, Discussions, Discussion/Conclusions, Discussions/Conclusions, Discussion and implications of the research, Discussion and perspectives, Discussion and recommendations, Discussion and summary, Discussion of Forum Themes, Discussion-Conclusion, Discussion/Conclusion, Discussions and conclusion
Conclusion	Conclusion, Conclusions, Conclusions/Significance, Conclusion/Significance, Conclusions and Significance, Conclusions/interpretation, Conclusion and Significance, Conclusions/ Significance, Conclusion(s), Conclusions / Significance, Conclusions and implications, Conclusions and recommendations, Conclusions/Signficance, Conclusions/Significances, Conclusion/ Significance, Conclusion and recommendations, Conclusions & Significance, Conclusions/Findings, Conclusions/Significance Abstract, Conclusion & Recommendation, Conclusion / Significance, Conclusion and clinical importance, Conclusion and implications

4.3. 섹션 맵핑

사용된 빈도수를 기준으로 상위 50개의 섹션명은 빈도수를 기준으로 전

체의 98.97%에 해당하였는데, 나머지 것들은 대부분 한번 사용된 경우이었기 때문이었다. 이 50개의 섹션명들을 IMRAD로 맵핑하는 것은 NLM에서 제공하는 섹션명 맵핑 파일(NLM 2012)을 사용하였다. 맵핑 파일은 PMC article에서 사용된 섹션명들을 “Objective”, “Background”, “Methods”, “Results”, 그리고 “Conclusion”중 하나로 맵핑한 리스트를 제공하는데, 본 연구에서는 맵핑된 결과를 IMRAD로 변환하기 위해 “Objective” 와 “Background”를 “Introduction”으로, “Conclusion”은 “Discussion”으로 맵핑하였다.

4.4. 언어적 특징 추출

구조화된 초록을 IMRAD로 맵핑한 후에는, 각 섹션의 언어적 특징을 추출하고 그 결과를 이용하여 섹션별 빈도 테이블을 구축하였다. 추출 대상은 각 문장에 사용된 동사/명사, 동사구/명사구, 그리고 n-gram이었고, 추출을 위한 도구로는 자연어 처리 분야에서 많이 사용되는 LingPipe[3]를 사용하였다. LingPipe는 아파치 재단(Apache Foundation)의 openNLP와 함께 자연어 처리 분야에서 좋은 성능으로 인해 자주 사용되는 소프트웨어이다(Kang, van Mulligen et al. 2011). 본 연구에서는 LingPipe의 품사 태거인 MedPost (Smith, Rindflesch et al. 2004)파서를 이용하여 문장 내 각 단어들의 품사 정보를 얻은 후, 이를 이용하여 구 (Phrase) 와 N-gram ($2 \leq n \leq 10$) 을 추출하였다. N-gram을 추

[3] http://alias-i.com/lingpipe/

출할 때는, 숫자를 사용한 표현을 일반화하기 위해서 “NUM” 심볼로 대체하여 처리하였다. 추출된 각 특징은 섹션 정보와 함께 빈도 정보를 갱신하여, 최종적으로 섹션에서 각 언어적 특징의 사용빈도를 얻었다. 이때, 빈도 테이블의 크기를 줄이기 위해 빈도수가 낮은 것들은 제외하였다.

추출된 언어적 특징들은 분류 시스템의 데이터로 사용되기 위하여 가중치를 계산하였는데, 사용한 수식은 아래와 같다. 수식은 정보검색 분야에서 문서에 포함된 키워드의 가중치를 계산하는데 사용되는 TF-IDF(Term Frequency - Inverse Document Frequency) 식을 변형한 것이다.

$$w_{t,s} = TF(t,s) \times IDF(t,S)$$

$$TF(t,s) = Normalized\, f(t,s)$$

$$IDF(t) = log(\frac{|S|}{\sum_{s \in S}(f(t,s)/Maxf(t,S))})$$

위 식에서, 언어적 특징 t가 섹션 s에서 가지는 가중치 $w_{t,s}$ 는 t가 s에서 가지는 중요도 *TF(t, s)*와 전체 IMRAD 섹션 S에서 가지는 구별력 *IDF(t)*의 곱으로 계산된다. *TF(t, s)*는 t가 s에서 사용된 빈도수 *f(t, s)*를 각 섹션별 빈도수의 합을 이용하여 보정한 값으로 계산하고, *IDF(t)*는 특정 섹션에 집중되어 사용된 언어적 특징이 섹션 전체에 골고루 사용된 것보다 높은 값을 가지도록 일반적인 IDF를 변형하여 설계하였다. IDF를 계산하는 식에서 |S|는 IMRAD를 구성하는 섹션 개수 (여기서는 4), 그리고

*Max f(t, S)*은 t 가 IMRAD 섹션에서 사용된 빈도수 중 최대 값이다. 본 연구에서는 동사(구), 명사(구), 그리고 N-gram ($2 \leq n \leq 10$) 에서 빈도수가 높은 것들을 각각 추출하여 가중치를 계산하였다. 가중치를 계산한 후에는, Social Network 분석 프로그램인 Pajek(Batagelj and Mrvar 2002)을 이용하여 추출된 언어적 특징들과 섹션들과의 관계를 분석하였다.

5. 결과

5.1. 섹션별 동사/동사구의 사용 특징

“Introduction” 섹션에서 가장 빈번하게 사용된 것은, 연구의 배경이나 목적을 위해 사용된 조동사나 To부정사 표현이었다. 동사의 경우 To 부정사의 “to”와 “are”, 그리고 완료형 표현에 사용되는 “have”, “been”이 가장 많이 사용되었다. 동사구의 경우는 보다 명확하게 사용 형태가 보였는데, “to determine”, “to evaluate”, 그리고 “to identify”와 같은 연구 목적을 위한 표현과 “has been”과 “is known”과 같이 연구 배경을 설명하기 위한 것들이 높은 빈도로 사용되었다. “Methods” 섹션에서 동사들은 과거형과 수동태 표현으로 주로 사용되었다. 또한 높은 빈도로 사용된 동사들을 보면 “perform”, “measure”, “assess”와 같이 Williams가 분류한 “Methods” 에 해당하는 동사들이었다(Williams 1996). “Results” 섹션은 “Methods” 섹션과 같이 과거형과 수동태 표현이 주로 사용되었지만, 동사의 경우는 조금 다른 사용 행태를 보여주었다. “Results” 섹션에서는 주로 실험 결과를 통한 발견을 설명하는 과정

에서, Williams의 “Observation”에 해당되는 “observe”, “find”, “identify”, 그리고 “detect”와 같은 동사가 사용되었다(Williams 1996).

마지막으로 “Discussion” 섹션에서는 조동사가 높은 빈도수로 사용되었으며, 특히 “could”나 “should”와 같은 조동사의 사용은 “Introduction” 섹션과 비교하여 눈에 띄는 차이였다. Table 2와 Table 3은 각 섹션에서 가장 많이 사용된 상위 25개의 동사/동사구 리스트이다.

Table 2 섹션별 최다 사용 빈도 동사 (상위 25개)

Introduction	Methods	Results	Discussion
to	were	was	to
are	was	were	be
have	to	to	are
been	Using	had	may
was	used	compared	was
has	performed	showed	can
be	assessed	found	have
can	included	be	were
may	analyzed	are	suggest
used	conducted	associated	should
associated	compared	Using	associated
were	evaluated	observed	could
determine	collected	increased	has
evaluate	determined	have	will
investigate	had	did	provide
known	examined	used	used
investigated	obtained	revealed	Using
assess	treated	reported	found
including	received	detected	indicate
identify	underwent	decreased	might
reported	Including	show	suggests
shown	calculated	Including	provides

compared	aged	expressed	show
aimed	be	correlated	been
could	identified	could	compared

Table 3 섹션별 최다 사용 빈도 동사구 (상위 25개)

Introduction	Methods	Results	Discussion
to determine	was performed	was observed	may be
has been	were collected	was associated	can be
to evaluate	were measured	was found	is associated
is known	were compared	were identified	could be
to identify	was conducted	were found	should be
To assess	were analyzed	were observed	to be
to investigate	was assessed	were associated	may have
is associated	was used	was detected	will be
was to evaluate	were included	were detected	to improve
was to investigate	were used	to be	might be
was to determine	were assessed	were included	can be used
may be	were evaluated	was seen	was associated
have been	was measured	to identify	to identify
was to assess	were performed	was found to be	should be considered
To study	were obtained	was identified	appears to be
To compare	were determined	were obtained	to reduce
have shown	were treated	was performed	may provide
to examine	was determined	was increased	may contribute
are associated	was evaluated	was reduced	are associated
can be	were examined	were found to be	may play
to improve	was carried	were used	would be
was to examine	were enrolled	was used	To prevent
is unknown	were divided	were analyzed	will provide
To understand	To assess	were treated	is required
to develop	were calculated	to have	were associated

5.2. 섹션별 N-gram의 사용 특징

N-gram은 n의 크기를 2에서 10까지 증가시키면서 추출하였는데, n의 값이 3이상부터 문장의 목적을 나타내는 것들이 발견되었다. 예를 들어 연

구 목적을 위해 사용된 “study was to”나 “of this study was”와 같은 것들은 “Introduction” 섹션에서 가장 많이 사용된 것들이며, 연구 성과를 표현할 때 사용되는 “Results suggest that”, “this is the first”, 또는 “The results of this study”와 같은 표현은 “Discussion” 섹션에서 가장 많이 사용되었다.

“Methods”와 “Results” 섹션의 경우는 의생명 분야의 특성상 통계적 유의미성을 표현하기 위한 N-gram들이 눈에 띄게 많이 사용되었다. 예를 들어, “Results”섹션에서는 p value를 이용한 *‘p = NUM’* 나 *‘p < NUM’*과 같은 표현들이 대표적인 사례이었다. 그리고 “Methods”섹션에서는 비록 높은 순위는 아니지만 “were divided into NUM groups”나 “A total of NUM patients”같은 대상 환자에 대한 표현이 사용되었다.

Table 4 주요 n-gram (3 ≤ n ≤ 6, 상위 5개)

N-gram (N)	Introduction	Methods	Results	Discussion
n=3	study was to	NUM, NUM	(NUM %	as well as
	of this study	n = NUM	NUM %)	Results suggest that
	this study was	= NUM)	NUM) .	in patients with
	The aim of	(n =	NUM , NUM	the use of
	aim of this	NUM and NUM	= NUM)	the development of
n=4	of this study was	(n = NUM	(NUM %)	this is the first
	this study was to	n = NUM)	NUM (NUM %	these results suggest that
	The aim of this	NUM , NUM ,	p = NUM)	our results suggest that
	aim of this study	A total of NUM	= NUM) .	can be used to
	The purpose of	= NUM) .	p < NUM)	for the first

	this			time
n=5	of this study was to The aim of this study aim of this study was The purpose of this study purpose of this study was	(n = NUM) NUM , NUM , NUM , NUM , NUM , n = NUM) , NUM , NUM and NUM	NUM (NUM %) (p = NUM) (p < NUM) p = NUM) . p < NUM) .	The results of this study play an important role in this is the first report for the first time that our knowledge this is the
n=6	aim of this study was to The aim of this study was purpose of this study was to The purpose of this study was objective of this study was to	NUM , NUM , NUM , (n = NUM) , (n = NUM) . , NUM , NUM , NUM NUM , NUM , and NUM	NUM % CI NUM . NUM (p < NUM) . (p = NUM) . , NUM (NUM %) (P < NUM) .	our knowledge this is the first knowledge , this is the first the best of our knowledge , To the best of our knowledge the best of our knowledge ,

5.3. 섹션별 명사(구)의 사용 특징

명사와 명사구는 동사/동사구에서 보여준 섹션 구별력보다는 전반적으로 낮은 모습을 보여주었다. 코퍼스가 의생명 분야의 논문이란 점에서 “cancer”, “disease”, 그리고 “cell”과 같은 명사나 “gene express”과 “breast cancer”와 같은 명사구는 여러 섹션에서 높은 빈도로 사용되었다.

반면, 연구의 목표를 설명할 때 사용되는 “Present study”나 “our

aim"과 같은 명사구는 "Introduction"에서 많이 사용되었으며, "Discussion"섹션에서는 "our results", "our findings", 그리고 "our data"와 같은 표현이 최상위 빈도로 사용되었다. 또한 연구 방법과 관련된 "control group", "mean age", "confidence interval", 그리고 "odds ratio"들은 "Methods"와 "Results" 섹션에서 집중적으로 사용되었다. Table 5과 Table 6은 섹션별 최대 빈도의 명사와 명사구의 리스트이다.

Table 5 섹션별 최대 빈도 명사 (상위 25개)

Introduction	Methods	Results	Discussion
patients	patients	we	health
we	analysis	expression	patients
cancer	groups	analysis	trial
breast	treatment	levels	we
disease	expression	cell	treatment
treatment	cells	gene	intervention
clinicAl	cell	groups	clinicAl
risk	cancer	protein	research
cells	blood	treatment	data
care	Questionnaire	time	evidence
these	subjects	risk	risk
health	samples	while	their
studies	women	number	information
cell	time	activity	interventions
effects	levels	months	effectiveness
COMMON	mg	differences	patient
factors	hospital	cancer	quality
therapy	participants	health	studies
syndrome	weeks	ratio	knowledge
expression	risk	response	development
arthritis	factors	LESS	outcomes

patient	disease	factors	effects
mortality	period	subjects	factors
cases	sample	ml	disease
human	activity	children	approach

Table 6 섹션별 최대 빈도 명사구 (상위 25개)

Introduction	Methods	Results	Discussion
Present study	control group	confidence interval	our results
gene expression	mean age	significant difference	our findings
Breast cancer	cross-sectional study	significant differences	our data
Risk factors	consecutive patients	odds ratio	our study
important role	flow cytometry	control group	Further studies
previous studies	Blood samples	mean age	Present study
few studies	logistic regression	multivariate analysis	important role
physical activity	healthy controls	gene expression	first time
our study	body mass index	our results	further research
Recent studies	statistical analysis	significant increase	gene expression
Current study	medical records	T cells	Breast cancer
Cardiovascular disease	Risk factors	significant association	future studies
Oxidative stress	years old	hazard ratio	our knowledge
our aim	confidence intervals	cell lines	further investigation
Diabetes mellitus	Western Blot	higher levels	Risk factors
insulin resistance	retrospective study	significant correlation	physical activity
Metabolic syndrome	odds ratios	healthy controls	Oxidative stress
molecular mechanisms	physical activity	Risk factors	first report
colorectal cancer	prospective study	body mass index	useful tool
Recent years	blood pressure	high levels	High prevalence
Rheumatoid arthritis	Breast cancer	significant reduction	future research
Our objective	cross-sectional survey	Breast cancer	further study
primary care	case-control study	significant decrease	clinical practice

general population Prostate cancer	Focus groups group A	median age study period	clinical trials primary care

5.4. 언어적 특징들의 섹션 구별력

5.4.1. 동사

Table 7 은 각 섹션에서 가중치가 높은 동사를 10개씩 모아 정리한 것이며,

Figure 3은 섹션별 동사의 가중치 정보를 이용하여 SNA(Social Network Analysis)도구인 Pajek으로 분석한 결과 그래프이다. 그래프에서 각 노드(Vertex)는 동사 또는 섹션을 나타내며, 노드와 노드를 연결하는 선(edge)은 동사가 특정 섹션에 사용되었음을 나타낸다. 선의 굵기는 가중치에 의해 결정되는데, 선의 굵기가 두꺼우면 해당 동사와 섹션의 관련성이 높음을 나타내며, 반대의 경우는 관련성이 낮음을 나타낸다. 분석을 위해서 가중치가 낮은 선은 삭제한 서브 그래프를 사용하였다. 이와 같은 과정은 모든 분석에서 동일하게 이루어졌다.

Table 7 섹션별 주요 동사의 가중치(상위 10개)

Section	Verb	Weight			
		Introduction	Methods	Results	Discussion
Introduction	to	0.9704	0.3344	0.2028	0.8691
	been	0.6034	0.0325	0.0389	0.1350
	are	0.5399	0.0369	0.0952	0.5563
	has	0.5022	0.0126	0.0316	0.2367
	have	0.4488	0.0273	0.0717	0.3140
	was	0.4162	0.9097	0.7692	0.3988
	be	0.2962	0.0400	0.1072	0.9720

	may	0.2003	0.0043	0.0222	0.6990
	investigate	0.1411	0.0214	0.0067	0.0205
	can	0.1403	0.0058	0.0245	0.3016
Methods	were	0.1355	1.2264	0.6634	0.2570
	was	0.4162	0.9097	0.7692	0.3988
	Using	0.0181	0.3345	0.0886	0.1075
	to	0.9704	0.3344	0.2028	0.8691
	performed	0.0414	0.1740	0.0276	0.0311
	assessed	0.0385	0.1480	0.0160	0.0133
	used	0.0868	0.1385	0.0328	0.0906
	analyzed	0.0243	0.1237	0.0205	0.0087
	conducted	0.0445	0.1228	0.0085	0.0110
	compared	0.0710	0.1160	0.1563	0.1025
Results	was	0.4162	0.9097	0.7692	0.3988
	were	0.1355	1.2264	0.6634	0.2570
	had	0.0281	0.1034	0.2357	0.1020
	to	0.9704	0.3344	0.2028	0.8691
	showed	0.0204	0.0060	0.2023	0.1183
	found	0.0588	0.0090	0.1782	0.1516
	compared	0.0710	0.1160	0.1563	0.1025
	observed	0.0290	0.0153	0.1203	0.0969
	increased	0.0203	0.0033	0.1121	0.0535
	be	0.2962	0.0400	0.1072	0.9720
Discussion	be	0.2962	0.0400	0.1072	0.9720
	to	0.9704	0.3344	0.2028	0.8691
	may	0.2003	0.0043	0.0222	0.6990
	are	0.5399	0.0369	0.0952	0.5563
	suggest	0.0280	0.0006	0.0130	0.4018
	was	0.4162	0.9097	0.7692	0.3988
	should	0.0320	0.0022	0.0063	0.3851
	have	0.4488	0.0273	0.0717	0.3140
	can	0.1403	0.0058	0.0245	0.3016
	were	0.1355	1.2264	0.6634	0.2570

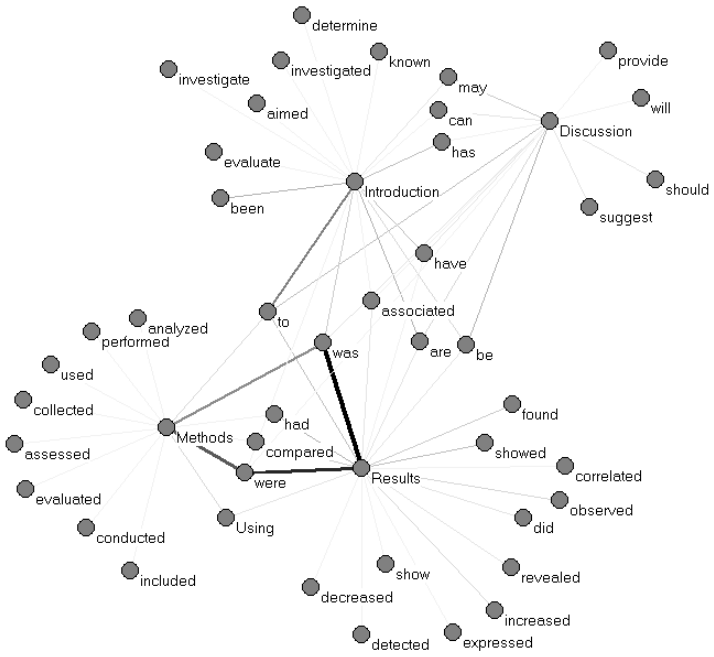

Figure 3 SNA를 통해 본 섹션별 주요 동사

동사의 경우, 선의 가중치 값이 가장 높은 “was”와 “were”가 “Results”와 “Methods”에서 사용되었음을 알 수 있었다. “was”와 “were”는 수동태 및 과거형 표현에 사용되는 동사로, 이러한 결과는 (Hanania and Akhtar 1985) 의 연구에서 언급한 과거 시제의 사용과 수동태의 사용이 "Results" 섹션에서도 나타남을 보여주었다. 또한 조동사의 사용이 "Discussion" 섹션뿐 아니라, "Introduction"에서도 많이 사용되었는데 이는 연구 배경을 설명하는 과정에서 사용된 완료형 표현으로 인한 것이었다. “Introduction”과 “Discussion” 섹션 사이에서는 주로 “may”, “has”, 그리고 “can”가 공통적으로 많이 사용되었으며, 연구 배

경 및 목표와 관련된 “aimed”, “been”, 그리고 “known”은 “Introduction” 섹션에서, 연구 성과의 의미와 기여를 설명하는데 사용된 “provide”, “should”, “suggest”, 그리고 “will”은 “Discussion” 섹션에서 주로 사용되었다. "Introduction" 섹션에서 가장 높은 가중치를 가진 것은 "to", "been", 그리고 "has" 였으며, "Discussion"에서는 "be", "may", 그리고 "should"이었다.

"Introduction"섹션과 "Discussion" 섹션에서 공통적으로 사용되는 동사가 많다면, "Methods" 섹션은 "Result" 섹션과 공통적으로 사용되는 것 많았다. 이 둘, 섹션에서 사용되는 동사는(Williams 1996)의 동사 분류처럼 섹션의 목적에 맞는 동사들이 사용되었으며, "compared"가 공통적으로 사용되었다. 또한 "associated"는 "Results" 섹션과 "Introduction"섹션에서 공통적으로 많이 사용됨을 알 수 있었다.

5.4.2. 동사구

Table 8은 각 섹션에서 가중치가 높은 동사구를 10개씩 모아 정리한 것이며, Figure 4 은 이 가중치 정보를 이용하여 구축한 그래프이다. 그래프에서 각 노드(Vertex)는 동사구 또는 섹션을 나타낸다.

Table 8 섹션별 주요 동사구의 가중치(상위 10개)

Section	Verb Phrase	Weight			
		Introduction	Methods	Results	Discussion
Introduction	is known	0.5656	0.0014	0.0047	0.0248
	has been	0.5141	0.0039	0.0237	0.1224
	was to evaluate	0.4484	0.0036	0.0010	0.0012

	was to investigate	0.4338	0.0025	0.0010	0.0000
	was to determine	0.4210	0.0057	0.0014	0.0003
	to determine	0.4105	0.1151	0.0450	0.1207
	to evaluate	0.4015	0.0865	0.0250	0.1231
	to investigate	0.3598	0.0562	0.0387	0.0827
	have been	0.3111	0.0052	0.0240	0.0749
	may be	0.3057	0.0030	0.0530	1.5828
Methods	was performed	0.0436	0.4557	0.0887	0.0165
	were measured	0.0173	0.3977	0.0415	0.0027
	were collected	0.0257	0.3892	0.0508	0.0031
	was conducted	0.0592	0.3681	0.0254	0.0081
	was assessed	0.0290	0.3472	0.0396	0.0042
	were assessed	0.0228	0.3382	0.0404	0.0026
	were compared	0.0420	0.3349	0.0646	0.0052
	were analyzed	0.0270	0.3303	0.0720	0.0059
	was measured	0.0104	0.3143	0.0331	0.0027
	were evaluated	0.0333	0.3136	0.0460	0.0028
Results	was observed	0.0113	0.0193	0.5324	0.1842
	was associated	0.0376	0.0059	0.4348	0.3521
	was found	0.0180	0.0070	0.4321	0.1714
	were found	0.0098	0.0088	0.3841	0.1183
	were observed	0.0086	0.0334	0.3678	0.1076
	were identified	0.0242	0.1406	0.3017	0.1056
	were associated	0.0299	0.0108	0.2127	0.1648
	was detected	0.0078	0.0371	0.1942	0.0430
	were detected	0.0066	0.0440	0.1881	0.0451
	was seen	0.0030	0.0015	0.1245	0.0358
Discussion	may be	0.3057	0.0030	0.0530	1.5828
	should be	0.0447	0.0020	0.0100	0.4717
	can be	0.1415	0.0027	0.0298	0.4196
	could be	0.0764	0.0031	0.0357	0.3792
	should be considered	0.0101	0.0003	0.0025	0.3791
	will be	0.0391	0.0243	0.0065	0.3698
	may have	0.0783	0.0009	0.0085	0.3527

	was associated	0.0376	0.0059	0.4348	0.3521
	is associated	0.2548	0.0026	0.0252	0.3499
	might be	0.0592	0.0017	0.0151	0.3458

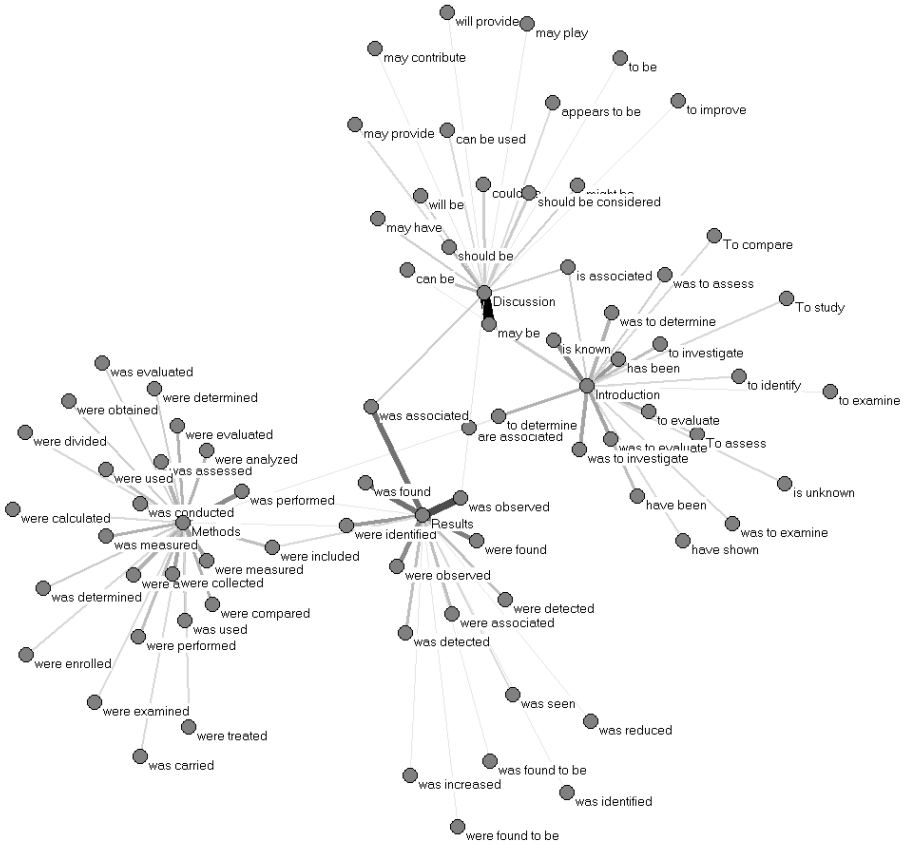

Figure 4 SNA를 통해 본 섹션별 주요 동사구

동사구의 경우는, 동사와는 달리 각 섹션의 특징을 나타내는 표현들이 잘 나타났다. 동사의 경우와 같이 "Results", "Methods"섹션에서는 과거 시제와 수동태 표현이 주로 사용되었으며, "Introduction"섹션과 "Discussion"섹션에서는 "may be"와 "can be"와 같은 조동사가 포함된

표현들이 공통적으로 섹션 구별력이 높았으며, "is associated"나 "are associated"와 같은 표현도 높은 구별력을 보여주었다. 또한 "to determine"은 "Introduction"섹션과 "Methods"섹션에서, "was associated"는 "Results"섹션과 "Discussion"섹션에서 높은 가중치를 보여주었다.

5.4.3. 명사

Table 9은 각 섹션에서 가중치가 높은 명사를 10개씩 모아 정리한 것이며, Figure 5은 이 가중치 정보를 이용하여 분석 결과 그래프이다. 그래프에서 각 노드(Vertex)는 명사 또는 섹션을 나타낸다.

Table 9 섹션별 주요 명사의 가중치 (상위 10개)

Section	Noun	Weight			
		Introduction	Methods	Results	Discussion
Introduction	patients	1.0112	1.1313	0.0000	0.8855
	we	0.9443	0.0273	0.5812	0.6695
	disease	0.3912	0.1240	0.0000	0.2026
	cancer	0.3376	0.1366	0.0790	0.1314
	care	0.3246	0.0001	0.0000	0.0017
	clinicAl	0.3195	0.0000	0.0760	0.4939
	cells	0.3129	0.2619	0.0000	0.0404
	these	0.3070	0.0000	0.0000	0.0017
	risk	0.2762	0.1284	0.1409	0.3019
	treatment	0.2677	0.1934	0.1546	0.4640
Methods	patients	1.0198	1.2234	0.0000	0.8946
	analysis	0.0473	0.3821	0.2238	0.0959
	cells	0.3156	0.2844	0.0000	0.0409
	expression	0.1794	0.2612	0.3648	0.0282
	groups	0.0267	0.2359	0.1934	0.0897

	women	0.1275	0.2279	0.0000	0.1698
	Questionnaire	0.0089	0.2276	0.0155	0.0241
	mg	0.0137	0.2193	0.0000	0.0141
	cell	0.1935	0.2153	0.2058	0.0327
	treatment	0.2701	0.2099	0.1560	0.4688
Results	we	0.9608	0.0299	0.6237	0.6842
	expression	0.1810	0.2632	0.3885	0.0285
	levels	0.1315	0.1929	0.2983	0.1070
	gene	0.1268	0.1410	0.2430	0.0262
	analysis	0.0477	0.3850	0.2384	0.0970
	cell	0.1953	0.2169	0.2193	0.0330
	protein	0.1179	0.1353	0.2151	0.0170
	groups	0.0270	0.2377	0.2061	0.0907
	months	0.0173	0.0000	0.1784	0.0454
	treatment	0.2725	0.2115	0.1662	0.4741
Discussion	health	0.2541	0.0122	0.1185	0.9464
	patients	1.0198	1.2234	0.0000	0.8946
	trial	0.0302	0.1001	0.0000	0.6833
	we	0.9524	0.0297	0.5861	0.6765
	clinicAl	0.3223	0.0000	0.0767	0.4991
	treatment	0.2701	0.2099	0.1560	0.4688
	their	0.0003	0.0001	0.0001	0.4127
	research	0.0566	0.0026	0.0184	0.3775
	risk	0.2786	0.1395	0.1421	0.3051
	evidence	0.0004	0.0197	0.0001	0.2936

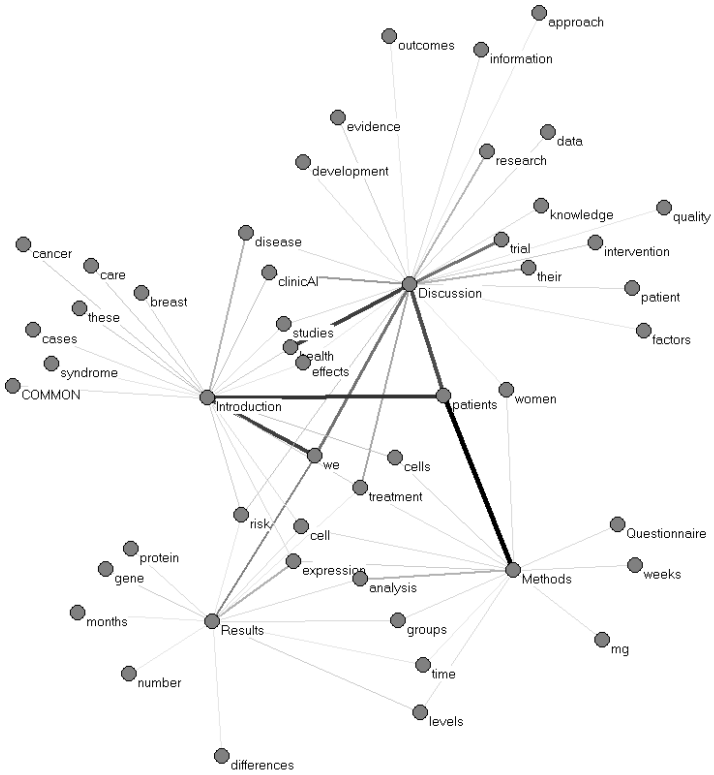

Figure 5 SNA를 통해 본 섹션별 주요 명사

명사의 사용은 동사나 동사구와는 달리, 섹션간에 비슷한 가중치를 가지는 것들이 많았다. 특히 "patients"와 "we"는 "Discussion", "Methods", 그리고 "Discussion"에서 비슷한 가중치를 가지고 보였으며, "disease", "health", 그리고 "effects"는 "Discussion"섹션과 "Introduction"섹션에서 비슷한 가중치를 보여주었다. 또한 "studies"는 "Discussion"섹션과 "Introduction"섹션에서 높은 가중치를 보여주었지만, "Discussion"섹션에서 가장 높은 가중치를 보여주었다. "Results"섹션과 "Methods" 섹션간에는 비슷한 가중치를 가지는 "groups", "time",

그리고 "level"과 같은 명사들이 많았다.

5.4.4. 명사구

Table 10은 각 섹션에서 가중치가 높은 명사구를 10개씩 모아 정리한 것이며, Figure 6은 이 가중치 정보를 이용하여 구축한 그래프이다. 그래프에서 각 노드(Vertex)는 명사구 또는 섹션을 나타낸다.

Table 10 섹션별 주요 명사구의 가중치 (상위10개)

Section	Noun Phrase	Weight			
		Introduction	Methods	Results	Discussion
Introduction	Present study	1.1669	0.1256	0.0606	0.4259
	previous studies	0.3736	0.0162	0.0274	0.1238
	few studies	0.3618	0.0013	0.0044	0.0174
	Recent studies	0.3182	0.0013	0.0026	0.0096
	important role	0.2776	0.0007	0.0158	0.2684
	our aim	0.2713	0.0061	0.0015	0.0040
	Our objective	0.2307	0.0036	0.0006	0.0019
	Recent years	0.2133	0.0018	0.0064	0.0271
	our study	0.2114	0.0379	0.0438	0.6102
	Breast cancer	0.2097	0.0774	0.0441	0.1376
Methods	cross-sectional study	0.0490	0.4910	0.0118	0.0107
	consecutive patients	0.0106	0.4131	0.0189	0.0024
	control group	0.0268	0.3974	0.2392	0.0387
	mean age	0.0072	0.3812	0.2169	0.0040
	Blood samples	0.0269	0.2874	0.0272	0.0120
	medical records	0.0142	0.2639	0.0089	0.0123
	flow cytometry	0.0185	0.2610	0.0426	0.0097
	logistic regression	0.0071	0.2483	0.0316	0.0064
	statistical analysis	0.0174	0.2091	0.0300	0.0178
	retrospective study	0.0555	0.1857	0.0042	0.0063
Results	confidence interval	0.0045	0.0703	0.5426	0.0088

	significant difference	0.0067	0.0184	0.4535	0.1061
	significant differences	0.0225	0.0374	0.4348	0.1186
	odds ratio	0.0056	0.0767	0.4069	0.0065
	control group	0.0268	0.3974	0.2392	0.0387
	mean age	0.0072	0.3812	0.2169	0.0040
	multivariate analysis	0.0054	0.0842	0.1975	0.0123
	significant increase	0.0148	0.0024	0.1724	0.0592
	our results	0.0167	0.0131	0.1650	1.9735
	significant association	0.0057	0.0027	0.1483	0.0767
Discussion	our results	0.0167	0.0131	0.1650	1.9735
	our findings	0.0092	0.0098	0.0323	1.0628
	our data	0.0081	0.0074	0.0618	0.9555
	our study	0.2118	0.0387	0.0438	0.6102
	Further studies	0.0141	0.0022	0.0095	0.5877
	Present study	1.1687	0.1283	0.0606	0.4259
	further research	0.0148	0.0016	0.0032	0.3956
	future studies	0.0131	0.0013	0.0011	0.3364
	further investigation	0.0185	0.0050	0.0083	0.2698
	important role	0.2780	0.0007	0.0158	0.2684

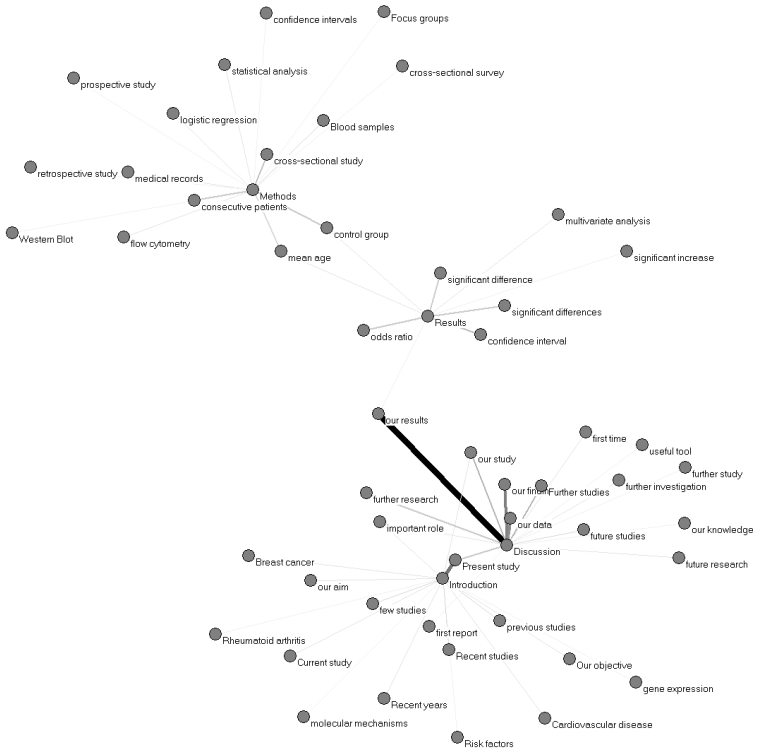

Figure 6 SNA를 통해 본 섹션별 주요 명사구

섹션별로 가중치가 높은 명사구들은 전체적으로 한 섹션에서만 높은 가중치를 가지는 모습을 보여주었다. 그러나, 연구의 목적이나 의미, 또는 기여를 설명을 언급할 때 사용되는 "Present study"와 "our study"등이 "Introduction" 섹션과 "Discussion"섹션에서 공통적으로 높은 가중치를 보여주었으며, 연구 결과를 언급하기 위한 "our results"는 "Discussion" 섹션과 "Results"섹션에서 주로 사용되었다. "Results"섹션과 "Methods"섹션간에는 "control group"과 "mean age"와 같은 명사구가 다른 섹션에 비해 상대적으로 높은 가중치를 보여주었다.

5.4.5. N-Gram

Table 11은 각 섹션에서 가중치가 높은 N-gram (n=3)를 10개씩 모아 정리한 것이며, Figure 7은 이 가중치 정보를 이용하여 구축한 그래프이다. 그래프에서 각 노드(Vertex)는 N-gram 또는 섹션을 나타낸다. 추출한 N-gram은 n의 크기를 2에서 10까지 증가시키면서 추출하였고, 본 장에서는 그중 n값이 3일때 가중치 테이블을 설명한다. 추출한 모든 N-gram에 대한 정보는 본 연구의 결과물을 설명한 웹 사이트인 http://abstract.bike.re.kr 에서 제공한다.

Table 11섹션별 주요 N-gram의 가중치 (상위10개, n=3)

Section	N-gram(n=3)	Weight			
		Introduction	Methods	Results	Discussion
Introduction	study was to	1.0050	0.0102	0.0007	0.0018
	of this study	0.8702	0.0161	0.0024	0.2832
	This study was	0.8528	0.0475	0.0015	0.0247
	the aim of	0.7060	0.0130	0.0010	0.0183
	aim of this	0.5389	0.0054	0.0003	0.0054
	The purpose of	0.3377	0.0100	0.0008	0.0169
	To evaluate the	0.3166	0.0811	0.0036	0.0739
	To investigate the	0.2889	0.0493	0.0046	0.0458
	purpose of this	0.2831	0.0040	0.0002	0.0048
	the present study	0.2782	0.0307	0.0061	0.2429
Methods	NUM and NUM	0.0844	0.5537	0.1877	0.1415
	n = NUM	0.0179	0.5073	0.0548	0.0051
	was used to	0.0266	0.3738	0.0125	0.0100
	NUM patients with	0.0128	0.3002	0.0206	0.0102
	NUM to NUM	0.0593	0.2833	0.1572	0.0875

	NUM NUM NUM	0.0370	0.2796	0.1267	0.0460
	were used to	0.0170	0.2741	0.0101	0.0064
	NUM NUM and	0.0284	0.2652	0.0997	0.0347
	NUM _ NUM	0.0038	0.2132	0.3110	0.0144
	a total of	0.0068	0.1852	0.0407	0.0072
Results	NUM NUM %	0.0064	0.0935	0.5594	0.0359
	P = NUM	0.0011	0.0075	0.4242	0.0165
	p < NUM	0.0011	0.0430	0.4167	0.0235
	NUM % CI	0.0015	0.0162	0.3236	0.0146
	NUM _ NUM	0.0038	0.2132	0.3110	0.0144
	NUM % of	0.1688	0.0747	0.2841	0.2804
	% CI NUM	0.0010	0.0027	0.2804	0.0130
	NUM % NUM	0.0121	0.0407	0.2382	0.0252
	and NUM %	0.0205	0.0816	0.2362	0.0550
	NUM % and	0.0186	0.0526	0.2340	0.0621
Discussion	Results suggest that	0.0008	0.0000	0.0036	0.7602
	is the first	0.0283	0.0012	0.0014	0.4020
	suggest that the	0.0105	0.0001	0.0030	0.3952
	we conclude that	0.0000	0.0000	0.0005	0.3868
	findings suggest that	0.0025	0.0000	0.0012	0.3735
	these results suggest	0.0005	0.0000	0.0013	0.3516
	results indicate that	0.0006	0.0001	0.0045	0.3503
	can be used	0.0467	0.0038	0.0042	0.3394
	our results suggest	0.0000	0.0000	0.0015	0.3368
	data suggest that	0.0061	0.0000	0.0017	0.3342

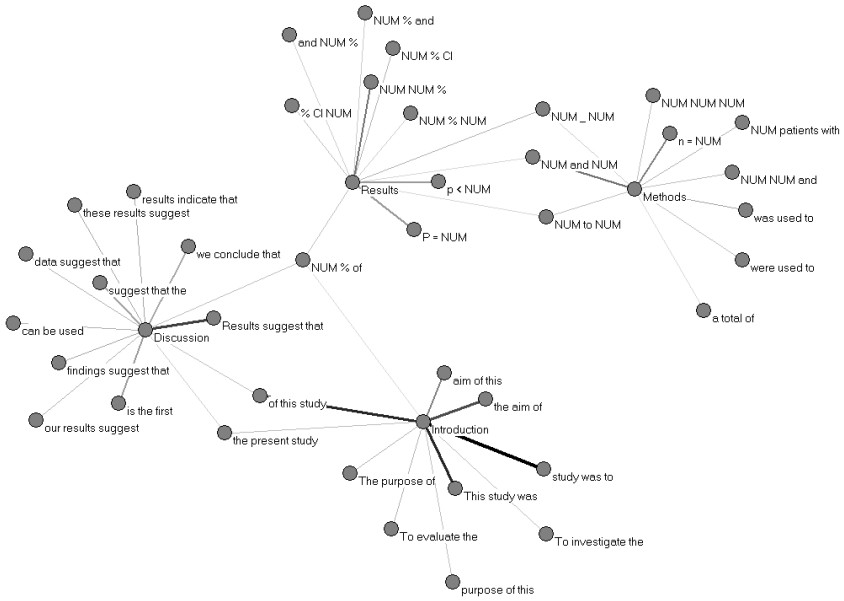

Figure 7 SNA를 통해 본 섹션별 주요 N-gram(n=3)

N-gram의 경우는 다른 언어적 특징과 유사하게 "Introduction"섹션과 "Discussion" 섹션에서 높은 구별력을 가지는 것들과, "Results"섹션과 "Methods"섹션에서 높은 구별력을 가지는 것들을 발견할 수 있었다. 이 중 첫 번째의 경우에 해당하는 것들은 "the present study"나 "of this study"등이며, 두 번째의 경우는 숫자와 관련된 표현인 "NUM and NUM", "NUM to NUM", 그리고 "NUM _ NUM"등이 있었다. "NUM % of"는 "Methods"를 제외한 나머지 섹션에서 비슷한 가중치를 보여주었다.

6. 결론

본 연구는 구조화된 초록을 구성하는 각 섹션의 언어적 특징을 추출하는 것을 목적으로, 152,083개의 구조화된 초록에서 동사(구), 명사(구),

그리고 N-gram ($2 \leq n \leq 10$)을 추출한 후 이들의 사용행태를 분석하였다. 분석과정에서는 각 언어적 특징의 섹션별 빈도수와 TFIDF기반의 가중치 공식을 이용하여 섹션에서의 중요도를 계산한 후, 이를 이용하여 네트워크 분석을 수행하였다. 코퍼스로 사용한 초록의 경우 다양한 이름의 섹션명이 사용되었으므로, 본 연구에서는 이를 IMRAD로 그룹화하는 작업을 수행하여 연구 결과물들이 비구조화된 초록을 IMRAD로 분류 하는데 활용될 수 있도록 하였다.

섹션별 사용 빈도수를 기준으로, 각 섹션의 언어적 특징은 섹션의 목적에 부합하는 행태를 보여주었다. "Introduction"은 연구 배경 및 연구 목적에 해당하는 문장들이 To 부정사와 완료형 표현들로 표현되었으며, "Methods"섹션의 경우는 연구 방법과 관련된 "perform", "measure", "collect", 그리고 "conduct"와 같은 동사들이 높은 구별력을 가지면서 사용되었고, 연구 결과의 관찰(Observation)과 보고(Reporting)에 해당하는 "show", "find"등은 "Results"섹션에서 주로 사용되었다. 이러한 동사들은 주로 과거시제와 수동태로 사용되었다. 이외에도 "Methods"와 "Results"섹션에서는 수치와 심볼을 사용한 표현들이 많이 사용되었는데, 이러한 행태는 사용된 코퍼스가 의생명 분야라는 특징을 잘 보여주었다. 특히 "Results"섹션에서 유의 확률 p값을 표현하는 N-gram들은 높은 가중치를 가지고 사용되었으며, 다른 섹션에서는 찾을 수 없는 특징이었다. 마지막으로 "Discussion" 섹션은 언어적 특징만으로는 "Introduction"과 유사한 점이 많았다. 조동사와 현재형의 사용이 주로 두 섹션의 공통점으로 나타났는데, "Discussion"에서는 "Introduction"와는 다른 "should",

"can", 그리고 "will"이 주로 사용되었다. 또한 "Discussion"섹션과 "Introduction"섹션은 섹션의 목적에 부합하는 명시적인 문형이 존재하였는데, 이러한 것들은 연구 목표를 위한 "study was to"나 "of this study", 연구 결과가 기여하는 바를 표현하는 "Results suggest that"등에서 확인할 수 있었다.

본 연구에서 추출한 언어적 특징들은 향후 의생명 분야의 비구조화된 초록을 IMRAD중 하나의 섹션으로 태깅하기 위한 데이터로 사용될 수 있다는 점에서 중요하며, 나아가 각 섹션의 특징을 잘 나타내는 문형 또는 패턴을 구축하는데 중요하게 사용될 수 있다.

Ⅲ. 언어적 특징을 이용한 초록 문장 분류

1. 연구 배경

1987년 구조화된 초록의 도입으로부터 의생명 분야에서는 논문의 유형에 따른 여러 형식의 들이 발표되어 왔는데(Huth 1987, Literature 1987, Group 1994, Nakayama, Hirai et al. 2005), 그 중에서도 초록을 IMRAD로 구분한 형식이 가장 많이 사용되어 왔다 (Huth 1987, Sollaci and Pereira 2004). 의생명 분야의 대표적인 서지 데이터베이스인 MEDLINE (Medical Literature Analysis and Retrieval System Online)의 구조화된 초록의 현황을 보면, 1989년에서 1991년까지는 0.4%에 지나지 않았지만 1992년에서 2006년에서는 13.1%까지 증가하였고 2008년에는 23.0%까지 이를 정도로 꾸준히 증가하고 있다(Ripple, Mork et al. 2011). 그러나, 이러한 증가에도 불구하고 MEDLINE에 해마다 추가되는 논문들 중 75%에 이르는 초록은 여전히 구조화된 형식을 따르지 않고 있다(Ripple, Mork et al. 2012).

구조화된 초록이 가지는 장점은 초록에서 정보를 추출하거나 검색을 위한 색인과정에서 잘 나타난다. 먼저 의생명 분야에서 대규모의 문헌에서 자동으로 정보를 추출하는 분야에서, 초록은 중요한 정보 소스이며 특히 구조화된 초록은 섹션정보를 포함하고 있어 더욱 정확한 정보 추출이 가능하다. 검색 분야에서도 구조화된 초록은 IMRAD 섹션 중 하나에 대해서 검색이 가능하도록 해, 연구자가 더욱 정확한 검색을 할 수 있도록 한다. 이러한 이유로 의생명 분야에서는 비구조화된 초록을 구조화하려는 다양

한 시도가 이루어지고 있다. 본 연구도 이러한 연구의 흐름에서 비구조화된 초록의 문장들을 IMRAD 로 자동으로 태깅하는 것을 목적으로 하고, 특히 의생명 분야의 언어적 특징을 활용하는 방법을 제안한다.

2. 연구 목적

본 연구의 목적은 의생명 분야의 비구조화된 초록을 구조화하는 것을 목적으로 하며, 이러한 목적을 달성하기 위해 비구조화된 초록의 문장들을 IMRAD 섹션 중 하나로 분류하는 시스템을 개발하는 것이다.

자동 분류 시스템의 성능 향상을 위해서, 본 연구에서는 II장에서 구축한 섹션별 가중치 테이블을 사용하며, 이를 통해 적은 계산 비용으로도 높은 분류 성능을 낼 수 있는 방법을 제안한다. 일반적으로 분류 성능은 어떤 분류 알고리즘을 사용했는가와 어떤 특징들을 사용했는가에 의해 결정되는데, 본 연구는 의생명 분야의 언어적 특징을 효율적으로 문장 분류의 특징에 반영하였다는 점에서 기존의 연구와 차이점이 있다.

3. 관련 연구

의생명 분야에서 초록의 문장을 자동으로 분류하는 연구는 지속적으로 이루어져왔다. 본 장에서는 분류 특징들 관점에서 관련 연구들을 소개한다.먼저, (Hirohata, Okazaki et al. 2008)는 초록의 문장을 분류하기 위해서 문장을 구성하는 각 단어, bigram, 문장의 위치, 그리고 앞 뒤 문장의 단어들과 bigram을 분류 특징으로 사용하였다. 이 연구에서 사용한 분류 알고리즘은 Conditional Random Fields(CRFs)을 사용하였으며,

51,000개의 구조화된 초록에서 작성한 두 가지 코퍼스를 대상으로 실험하였다. 그 결과, 문장 수준에서는 최대 95.5%의 Accuracy와 초록 단위로는 64%의 Accuracy를 보여주었다.

(Ruch, Boyer et al. 2007)의 연구는 초록에서 핵심 문장들을 추출하였는데, 이를 위해 어근 처리를 한 n-gram (1≤n≤ 3), 문장 길이, 그리고 문장의 상대적인 위치 정보를 분류 특징으로 사용하였다. 이 연구에서는 나이브 베이지안 분류 알고리즘을 사용하여 비구조화된 초록의 "Conclusion"섹션에 대해서 84%의 F-score을 보여주었다. (Guo, Korhonen et al. 2010)는 초록을 구성할 수 있는 3가지 정보 구조를 정의하고 이에 맞게 문장을 분류하는 연구 결과를 보여주었는데, 사용한 특징으로는 문장 위치, 문장을 구성하는 단어, bi-gram, 동사, 동사의 클래스, 품사(Part of Speech), 문장의 문법적 관계, 주어와 목적어, 그리고 태(Voice)를 사용하였다. 이 연구와 다른 연구와의 차이점은 문장에서 사용된 동사와 동사의 클래스를 분류 특징으로 사용하려 했다는 점이다. 그러나, 테스트 데이터가 1000개라는 점에서 동사와 동사의 클래스 정보가 의생명 분야의 특징을 충분히 반영할 수 있었는가에 대한 의문점이 생긴다. 이 연구에서는 Support Vector Machine(SVM)을 사용하여 세 가지의 테스트 코퍼스에 대해서 81% ~ 90%의 Accuracy를 보여주었다.

(McKnight and Srinivasan 2003)은 의생명 분야에서 문장 분류의 주 대상이 되는 Randomized Controlled Trials(RCTs)을 대상으로 실험을 하였다. 사용한 특징으로는 일반적으로 사용되는 Bag-of-words(BOW)와 문장 위치를 이용하였으며, SVM 분류 알고리즘을 이용하여 구조화된

초록과 비구조화된 초록에 대해서 테스트하였다. 이들의 방법은 주로 "Methods"와 "Results" 섹션에서 상대적으로 높은 F-score 값인 76.8%에서 85.1%의 성능을 보여주었다. (Yamamoto and Takagi 2005)는 PubMed의 검색 서비스를 개선시키기 위한 방안으로, 초록의 문장을 "background", "purpose", "method", "result", 또는 "conclusion"으로 분류하는 연구를 수행하였다. 이들은 다른 연구와 달리 다양한 언어적 특징을 사용하였다는 점에서 본 연구와 공통점이 많다. 이들이 사용한 특징으로는 문장 위치, TF/IDF 값, 조동사의 사용 여부, 동사의 시제, BOW, 그리고 문장에 포함된 단어와 bigram, 그리고 문장구조에 대한 카이제곱 값들을 사용하였다. 그러나 이러한 방법은 앞에서 설명한 (McKnight and Srinivasan 2003)과 비슷한 성능을 보여주었는데, 비구조화된 초록에서는 16.6%에서 81.0%까지의 섹션별 F-score를 보여주었고, 구조화된 초록에서는 45.2%에서 72.3%의 F-score를 보여주었다.

(Xu, Supekar et al. 2006)은 비구조화된 RCT 초록 문장을 분류하기 위한 연구로서, 특징으로는 BOW와 문장 위치를 사용하였다. 문장 분류 알고리즘으로는 Hidden Markov Model(HMM)을 나이브 베이지안과 Maximum Entropy(ME), 그리고 Decision Tree와 합친 새로운 방법을 사용하였다. 이러한 방법을 사용하여 구조화된 RCT 초록을 대상으로 사용하였다. 성능을 살펴보면, HMM과 ME를 합친 방법에서 85.7%에서 98.2%에 이르는 F-score를 보여주었다. (Lin, Karakos et al. 2006)은 구조화된 초록의 섹션들이 변경되는 것을 포착하기 위한 HMM기반의 내용 모델을 제안하였으며, 구조화된 초록과 비구조화된 초록에서 각각

78.1%에서 86.7%, 74.3%에서 88.5%의 F-score값을 보여주었다. 마지막으로 (Chung 2009)은 RCT 초록에서 핵심 문장을 추출하기 위한 방안으로 문장을 구성하는 단어, POS, 문장위치, 그리고 이전 문장에서 추출한 특징들을 사용하였다. 문장 분류 알고리즘으로는 CRF를 사용하여 비구조화된 초록과 구조화된 초록을 대상으로 93%에서 98%의 F-score를 보여주었다.

이상으로, 문장 분류와 관련된 연구들을 살펴보면 공통적으로 BOW와 문장 위치가 가장 빈번하게 사용되었다. 특히 초록에서는 문장의 위치가 문장을 분류하는데 중요한 정보를 제공함을 알 수 있었다. 반면, BOW나 bi-gram을 사용하는 경우는 문장 분류를 위한 벡터의 크기를 증가시켜 계산비용에 영향을 줄 수 밖에 없었다. 또한, 동사의 시제나 동사를 특징을 사용하는 경우도 있었지만, 이러한 언어적 특징이 의생명 분야의 특징을 잘 반영할 만큼의 규모에서 추출되지 않았을 뿐 아니라 언어적 특징을 사용하였더라도 어떤 언어적 특징이 의생명 분야의 초록에 있었는지에 대한 분석이 없었다. 본 연구는 대규모의 의생명 분야 구조화된 초록에 나타나는 언어적 특징을 분석하고 이를 문장 분류에 반영하였다는 점과 분석 결과를 이용하여 적은 계산 비용으로도 높은 수준의 문장 분류를 가능하게 하는 방법을 제안하였다는 점에서 기존 연구와 차별된다.

4. 연구 방법

4.1. Feature Set 구성

문장 분류에 사용하는 특징은 언어적 특징을 포함해서 총 4개의 그룹으

로 구성하였다.

언어적 특징 그룹

언어적 특징 그룹은 2장에서 구축한 N-gram, 명사(구), 그리고 동사(구)에 대한 가중치 테이블을 이용하였다. 문장에서 언어적 특징을 추출하기 위해서는

Figure 8 에서 제시한 알고리즘을 사용하였다. 알고리즘에서는 문장을 입력으로 받아, 가장 먼저 문장에서 n-gram을 포함한 3가지 언어적 특징을 추출한다. 그리고, 추출된 각각의 특징에 대해서 가중치 테이블을 이용하여 가중치 벡터 (weightVector) 를 가져오고, 이 값을 sumOfWeightVector 에 더한다. 모든 특징에 대해서 이러한 과정을 거친 후, 마지막으로는 sumOfWeightVector 값을 이용하여 섹션별 랭크 순위 벡터를 계산하여 반환한다. 예를 들어 sumOfWeightVector값이 [20.25, 0.25, 1.2, 12.3] 이라면, 반환되는 랭크 순위 벡터는 [1, 4, 3, 2]을 반환한다. 랭크 순위 벡터는 문장에서 추출한 언어적 특징들이 어떤 섹션을 가장 잘 나타내는지를 나타낸다.

Algorithm getRankOrderVector (S, T)
Input: String of sentence **S**, Lexical type **T** = [*Verb Phrase* / *N-gram* / *Noun Phrase*]
Output: Rank order vector **ROV**

1: Initialize S = 4 // size of IMRAD sections
2: Initialize weightVector[S]
3: Initialize sumOfWeightVector[S]

4: lexicalFeatures ← ExtractLexicalFeature(S, T)

```
5:  FOR EACH feature f IN lexicalFeatures {
6:    weightVector = getWeightVector(f) // using weighting tables
7:        sumOfWeightVector + = weightVector;
8:   }
9:  RETURN rank(sumOfWeightVector)
```

Figure 8 언어적 특징 구축 알고리즘

구축 알고리즘을 통한 3가지 특징 이외에, 조동사의 사용여부, 동사의 시제, 그리고 to 부정사의 사용여부에 대한 정보도 추출하여 사용하였다. 동사의 시제는 과거(PAST), 현재(PRESENT), 그리고 완료형(PERFECT)으로 구분하였으며, 나머지 값들은 불리언 값으로 표현하였다.

문법적 특징 그룹

문법적 특징 그룹에는 문장에서 사용된 품사 정보를 사용하였다. 품사 정보는 MedPost 태그 집합의 63가지의 품사 태그를 사용하였다.

구조적 특징 그룹

구조적 특징 그룹에는 문장의 위치와 이전 문장의 분류 결과를 사용하였다. 문장 위치는 초록의 문장 분류에 많이 사용되어온 특징으로, 본 연구에서는 초록의 평균 문장 수를 고려하여 1부터 20까지의 값으로 일반화 시켰다. 이전 문장의 분류 결과는 처음 문장을 고려하여, Start(1), Introduction(2), Methods(3), Results(4), 그리고 Discussion(5)중에 하나의 값을 가지도록 하였다.

Bag-of-words

Bag-of-words (BOW)는 문장 분류에서 가장 많이 사용되는 특징이다. 본 연구에서는 어근 처리를 하지 않은 상태로 문장의 각 단어로 구성하였다. 또한 현재 문장 이외에 앞 뒤 두 문장까지를 포함(Window = 2)하여 BOW 를 구성하였다. 중복된 단어가 발생하는 경우는, 구분을 위해서 접두사인 “PREV_”와 “POST_”를 단어에 각각 붙여주었다. 이렇게 앞뒤 두 문장을 추가한 것은 실험을 통해 결정되었는데, BOW의 분류 성능이 앞뒤 두 문장을 사용했을 때 가장 좋았기 때문이다.

4개의 특징 그룹은 본 연구에서 사용한 WEKA(Hall, Frank et al. 2009)의 타입 체계에 따라 Table 12과 같이 사용하였다. WEKA는 뉴질랜드의 Waikato 대학에서 개발된 기계학습 소프트웨어로, 자바 언어를 이용하여 쉽게 연동할 수 있는 장점으로 많은 연구 프로젝트에서 사용되고 있다.

Table 12 특징의 종류와 WEKA 타입

Category	Feature		Type of Value (WEKA)
Bag of words	Word token		string
Lexical Features	Verb Phrase	Introduction	nominal: 0,1,2,3,4
		Methods	nominal: 0,1,2,3,4
		Results	nominal: 0,1,2,3,4
		Discussion	nominal: 0,1,2,3,4
	Noun Phrase	Introduction	nominal: 0,1,2,3,4
		Methods	nominal: 0,1,2,3,4
		Results	nominal: 0,1,2,3,4
		Discussion	nominal: 0,1,2,3,4

	N-gram	Introduction	nominal: 0,1,2,3,4
		Methods	nominal: 0,1,2,3,4
		Results	nominal: 0,1,2,3,4
		Discussion	nominal: 0,1,2,3,4
	Verb Tense		nominal : PRESENT, PAST, PERFECT
	Modal verb		nominal :Yes | No
	To-Infinitive		nominal : Yes | No
Grammatical Features	Part of Speech		String
Structural Features	Sentence Location		nominal :1,...,20
	Sentence History		nominal : start, introduction, methods, results, discussion

4.2. 테스트 문서 집합

분류 시스템의 평가를 위해 3가지 테스트 문서 집합를 구축하였다. 첫 번째(SA)는 학습 및 평가를 위해 구조화된 초록으로 구성된 것이고, 두 번째(UA-1)와 세 번째(UA-2)는 비구조화된 초록을 수작업으로 태깅하여 구축하였다. SA는 152,083개의 구조화된 초록에서 랜덤하게 2,000개를 선택한 것으로, 총 23,881개의 문장으로 구성되었다. UA-1는 200개의 비구조화된 RCT 초록으로 구성되었으며, 총 1,786개의 문장으로 이루어졌다. RCT 초록은 PUBMED에서 논문의 타입을 RCT로 하여 검색한 결과에서 랜덤하게 선택하였다. UA-2는 200개의 비구조화된 초록으로 구성되었으며, 총 2,429개의 문장으로 이루어졌다. UA-2는 일반적인 비구조화된 초록에서 성능을 평가하기 위한 것으로, 2장에서 사용한 코퍼스의 비구조화된 초록 221,261개 중에서 랜덤하게 선택하였다.

UA-1과 UA-2는 의생명 분야 박사과정 2명에 의해서 모든 문장을 IMRAD중 하나로 태깅되었다. 태깅 과정 중에서 모호성이 발생하는 경우에는, 문장의 주절에서 전달하려는 메시지가 IMRAD의 어느 섹션에 가까운지를 기준으로 결정하였으며, 결정이 애매한 경우에는 두 어노테이터간의 토의에 의해 최종 섹션명이 결정되었다. 두 어노테이터 간의 동의 수준을 나타내는 Cohen's Kappa값은 UA-1과 UA-2에서 각각 0.913과 0.856이었다. 값으로 볼 때 RCT 초록이 일반적인 초록보다 각 문장의 목적이 명확함을 알 수 있었다. Table 13은 테스트 문서 집합 각각의 섹션 분포를 정리한 것이다.

Table 13 테스트 문서 집합의 섹션 분포

Dataset	Introduction	Methods	Results	Discussion
Structured Abstract Dataset (SA, n=2,000)	5,097 (21.3%)	6,244 (26.1%)	8,527 (35.7%)	4,013 (16.8%)
Unstructured Abstract Dataset-1 (UA-1, n=200)	479 (26.8%)	527 (29.5%)	531 (29.7%)	249 (13.9%)
Unstructured Abstract Dataset-2 (UA-2, n=200)	662 (27.3%)	274 (11.3%)	1,128 (46.4%)	365 (15.0%)

4.3. SVM을 이용한 학습 및 평가

분류 알고리즘은 Support Vector Machine(SVM)을 사용하였으며, polynomial kernel을 이용하였다. 평가는 먼저 테스트 문서 집합 SA를 대상으로 10-fold cross validation으로 하였다. 그리고 SA를 이용하여

구축한 모델을 사용하여 테스트 문서 집합 UA-1과 UA-2에 대해서 평가하였다. 섹션별 분류 성능은 Precision, Recall, 그리고 F-score를 사용하였으며, 전체 성능은 Accuracy를 사용하였다.

$$Precision = \frac{True\ Positive}{True\ Positive + False\ Positive}$$

$$Recall = \frac{True\ Positive}{True\ Positive + False\ Negative}$$

$$F - score = 2 \times \frac{Precision \times Recall}{Precision + Recall}$$

$$Accuracy = \frac{True\ Positive + True\ Negative}{Positive + Negative}$$

5. 연구 결과

5.1. 언어적 특징별 성능

Table 14 은 테스트 문서 집합 SA를 대상으로 하여 언어적 특징의 분류 성능을 실험한 결과이다. 6가지의 특징 중에서 N-gram이 71.22%의 Accuracy로 가장 좋은 성능을 보여주었다. 그리고 조동사의 사용여부만 사용한 경우가 가장 낮은 성능인 39.78% Accuracy를 보여주었다. 그리고 동사구는 N-gram보다는 낮은 성능이지만 모든 섹션에서 N-gram과 비슷한 성능을 보여주어, 동사구가 문장을 분류하는데 중요한 특징으로 사용될 수 있음을 보여주었다. 이러한 동사구의 성능은 명사구와 비교했

을 때 더욱 확실한데, 명사구는 45.97%의 Accuracy만을 보여주었기 때문이다.

Table 14 언어적 특징의 분류 성능

Feature	Accuracy	F-score			
		Introduction	Methods	Results	Discussion
N-gram	71.22	70.6	72.8	76.5	53.00
Verb Phrase	68.41	65.7	71.6	73.6	52.4
Noun Phrase	45.97	40.8	31.0	57.5	19.4
Verb Tense	46.27	54.4	-	61.4	-
Modal Verb	39.78	-	-	54.7	39.7
To Infinitive	40.24	37.0	-	55.9	-

Table 15은 언어적 특징 중 가장 높은 성능을 보여준 N-gram에 대해서 n의 크기를 증가시키면서 실험한 결과이다. 결과를 보면 n의 크기가 6이 될 때까지 성능이 향상되었고, 그 이후로는 거의 같은 성능을 보여주었다. 섹션별 결과를 보면 Methods 와 Results 섹션에서 n의 크기가 커짐에 따라 분류 성능이 좋아졌으며, 나머지 두 섹션은 n의 증가에 영향을 받지 않았다.

Table 15 N-gram의 성능

N-gram	Accuracy	F-score(precision/recall)			
		Introduction	Methods	Results	Discussion
N=2	68.80	71.3	68.7	72.7	52.9
N<=3	69.84	70.5	69.5	75.1	52.4
N<=4	71.11	70.6	72.5	76.4	52.9
N<=5	71.15	70.5	72.7	76.5	52.8
N<=6	71.24	70.6	72.8	76.6	53.0

N<=7	71.22	70.6	72.8	76.5	53.0
N<=8	71.22	70.6	72.8	76.5	53.0
N<=9	71.22	70.6	72.8	76.5	53.0
N<=10	71.22	70.6	72.8	76.5	53.0

5.2. 특징 그룹 조합별 성능

테스트 문서 집합 SA에서의 분류 성능

Figure 9 와 Table 16는 테스트 문서 집합 SA에 대해서 네 종류의 특징 그룹을 조합하여 실험한 결과이다. 먼저 네 종류의 특징 그룹 중 가장 좋은 성능을 보인 것인 BOW로 82.90%의 Accuracy를 보였으며, 구조적 특징 그룹의 경우도 단 2 차원의 특징 벡터만으로도 76.9%의 Accuracy를 기록하였다. 언어적 특징 역시도 15 차원의 특징 벡터로 76.3%의 Accuracy를 보여주었다. 구조적 특징 그룹과 언어적 특징 그룹을 살펴보면, 구조적 특징그룹의 경우는 Introduction 섹션에서 언어적 특징그룹보다 높은 F-score를 보여주었으나, Methods와 Results 섹션에서는 언어적 특징 그룹이 높은 F-score를 보여주었다. 언어적 특징 그룹은 Methods 섹션에서 가장 높은 성능을 보여주었는데, 이러한 성능은 네 종류의 특징 그룹 중에서 가장 높은 것이기도 하다.

두 가지 그룹을 조합한 경우에는 구조적 특징 그룹과 언어적 특징 그룹을 조합한 SL이 가장 높은 성능인 90.3% Accuracy를 보여주었다. 이러한 성능은 BOW를 사용한 조합보다 더 좋은 성능을 보였다는 점에서, 언어적 특징 그룹이 BOW를 대신할 수 있음을 보여주었다. 또한 BOW를 사용하는 방식은 구성상 특징 벡터의 차원이 높아져 계산 비용이 높아지게

되는데, 언어적 특징 그룹의 경우는 낮은 차원의 특징 벡터임에도 높은 성능을 보여주었다. 세가지 그룹을 조합한 경우에도 BOW를 제외한 GSL 조합이 75 차원의 특징 벡터만으로도 91.0%의Accuracy를 보여주었다. 마지막으로 모든 그룹을 사용한 BGSL은 91.7%의Accuracy를 보여주었다.

BOW를 사용하지 않으면서도 높은 성능을 보여준 SL과 GSL의 성능은, 모든 특징 그룹이 사용된 BGSL과 비교하여 Discussion섹션을 제외한 나머지 섹션에서 비슷한 수준의 성능을 보여주었다.

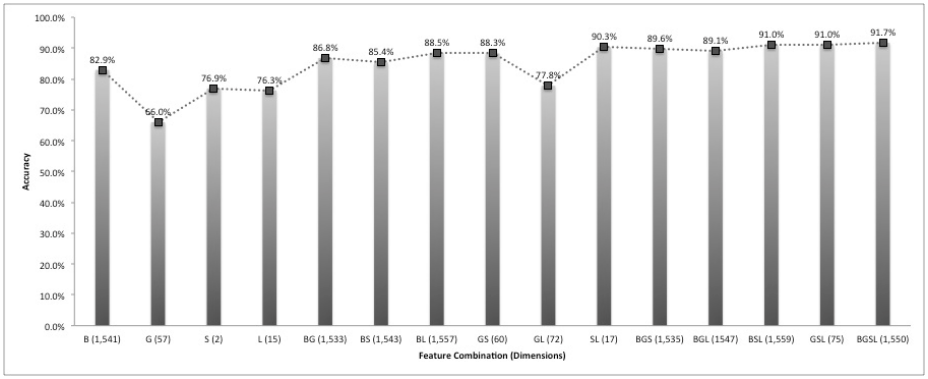

Figure 9 테스트 문서 집합 SA에서의 분류 성능 (Accuracy)

Table 16 테스트 문서 집합 SA에서의 분류 성능

Feature Combination (Dimensions)	Accuracy	F-score (Precision/Recall)			
		Introduction	Methods	Results	Discussion
B (1,541)	82.9	87.2 (85.9/88.6)	78.7 (78.8/78.5)	83.1 (83.2/83.0)	83.6 (84.9/82.3)
G (57)	66.0	64.2 (58.2/71.6)	69.2 (66.6/72.0)	72.6 (72.8/72.5)	45.0 (60.4/35.9)
S (2)	76.9	85.0 (81.1/89.2)	73.7 (68.8/79.3)	77.6 (76.1/79.0)	68.4 (95.7/53.2)
L (15)	76.3	74.4	81.0	80.7	60.2

		(71.8/77.1)	(82.3/79.8)	(78.1/83.6)	(67.4/54.3)
BG (1,533)	86.8	91.5 (90.8/92.3)	84.2 (83.9/84.6)	86.1 (86.1/86.1)	86.4 (88.0/84.8)
BS (1,543)	85.4	90.6 (90.1/91.0)	81.4 (80.5/82.4)	85.1 (85.4/84.9)	85.4 (87.0/83.9)
BL (1,557)	88.5	92.0 (91.3/92.7)	87.4 (86.9/87.8)	88.1 (88.0/88.2)	86.4 (88.2/84.7)
GS (60)	88.3	94.7 (94.4/95.0)	86.7 (83.4/90.2)	86.6 (88.0/85.3)	86.4 (89.8/83.3)
GL (72)	77.8	75.3 (73.6/77.1)	83.0 (83.9/82.2)	82.3 (79.6/85.2)	61.4 (68.1/55.9)
SL (17)	**90.3**	95.3 (95.1/95.5)	90.1 (88.6/91.6)	89.4 (88.8/89.8)	86.3 (90.3/82.7)
BGS (1,535)	89.6	94.3 (94.0/94.7)	87.3 (86.4/88.3)	88.8 (89.0/88.6)	88.7 (90.3/87.2)
BGL (1547)	89.1	92.4 (91.7/93.1)	87.9 (87.4/88.4)	88.8 (88.7/88.8)	87.5 (89.3/85.8)
BSL (1,559)	91.0	95.1 (94.9/95.3)	89.5 (89.1/89.9)	90.4 (90.0/90.8)	89.3 (91.2/87.6)
GSL (75)	91.0	95.9 (95.7/96.0)	91.0 (89.6/92.4)	90.0 (89.4/90.5)	87.1 (90.8/83.6)
BGSL (1,550)	**91.7**	95.8 (95.7/96.0)	90.4 (90.0/90.09)	91.0 (90.7/91.2)	90.2 (91.7/88.8)

테스트 문서 집합 UA-1에서의 분류 성능

Figure 10 과 Table 17는 테스트 문서 집합 UA-1에 대해서 네 종류의 특징 그룹을 조합하여 실험한 결과이다. 특징 그룹 중에서 가장 높은 Accuracy를 보인 것은, SA와 달리 언어적 특징 그룹이 77.3%을 기록하였다. 반면, SA에서 높은 성능을 보여준 구조적 특징 그룹은 가장 낮은 50.4%의 Accuracy를 보여주었다. 언어적 특징 그룹은 구조화된 초록으로 구성된 SA에서와 비슷한 수준의 높은 성능을 보여주었다.

두 개의 그룹을 조합한 경우에는, 구조적 특징과 언어적 특징을 조합했을 때 가장 높은 86.2%의 Accuracy를 보여주었는데, Methods 섹션과

Results 섹션에서 높은 성능을 보이는 언어적 특징과 Introduction 섹션과 Discussion 섹션에서 좋은 성능을 보여준 구조적 특징이 조합되면서 성능이 개선되었다. 세 개의 특징을 조합한 경우에도 SA와 같이 GSL이 86.2%의 Accuracy로 가장 높은 성능을 보여주었다. 이러한 성능은 BOW를 사용한 BSL보다 3.3%나 높은 값이며, 더욱 흥미로운 점은 모든 특징 그룹을 사용한 BGSL의 84.0% Accuracy보다 높은 성능이라는 것이다.

언어적 특징은 다른 특징 그룹과 사용될 경우, 눈에 띄는 성능 개선을 보여주었다. BOW(accuracy = 74.7%) 와 같이 사용된 경우인 BL에서는 82.4%의Accuracy를 보이며 7.7%의 성능향상을 기록하였다. 또한 구조적 특징 그룹(Accuracy= 50.4%)과 같이 사용된 경우인 SL에서는 86.2%의Accuracy를 보이며 35.8%의 성능향상을 보여주었다.

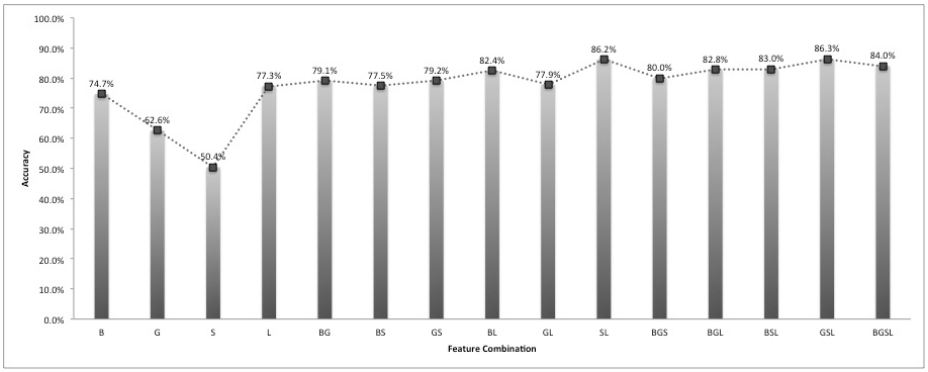

Figure 10 테스트 문서 집합 UA-1에서의 분류 성능 (Accuracy)

Table 17 테스트 문서 집합 UA-1에서의 분류 성능

Feature	Accuracy	F-Score (Precision/Recall)

Combination		Introduction	Methods	Results	Discussion
B	74.7	80.8 (85.0/77.0)	71.50 (73.4/69.6)	72.60 (71.4/73.8)	75.0 (68.3/83.1)
G	62.6	64.5 (64.964.1)	62.4 (65.2/59.8)	67.3 (61.0/75.1)	45.6 (55.1/39.0)
S	50.4	75.7 (68.1/85.2)	48.5 (39.0/64.1)	35.6 (47.6/28.4)	2.4 (100.0/1.2)
L	77.3	74.3 (80.0/69.3)	80.9 (82.2/79.7)	81.4 (76.4/87.2)	66.0 (65.4/66.7)
BG	79.1	84.3 (88.2/80.8)	75.8 (77.6/74.2)	77.4 (75.8/79.1)	79.6 (74.0/85.9)
BS	77.5	84.9 (88.3/81.8)	73.9 (70.1/78.0)	75.6 (75.7/75.5)	75.8 (79.6/72.3)
GS	79.2	90.1 (88.5/91.6)	75.0 (71.4/78.9)	74.2 (74.3/74.0)	77.8 (93.3/66.7)
BL	82.4	85.5 (91.6/80.2)	80.5 (80.1/80.8)	82.1 (82.0/82.3)	81.7 (74.5/90.4)
GL	77.9	75.5 (82.6/69.5)	79.9 (81.3/78.6)	82.3 (76.9/88.5)	68.4 (66.5/70.3)
SL	86.2	92.1 (93.2/91.0)	84.5 (82.0/87.1)	83.9 (82.4/85.5)	83.8 (92.3/76.7)
BGS	80	87.1 (91.1/83.5)	75.8 (72.5/79.5)	78.4 (78.0/78.7)	79.7 (82.4/77.1)
BGL	82.8	86.5 (91.0/82.5)	82.1 (83.1/81.2)	81.3 (81.5/81.2)	80.4 (72.7/90.0)
BSL	83.0	89.3 (94.4/84.8)	80.8 (77.6/84.3)	80.9 (81.2/80.6)	80.6 (79.4/81.9)
GSL	86.3	92.0 (93.1/90.8)	85.1 (82.2/88.2)	83.9 (83.0/84.7)	83.5 (91.0/77.1)
BGSL	84.0	89.5 (95.5/84.1)	82.0 (78.7/85.6)	82.2 (81.9/82.5)	82.8 (81.6/83.9)

테스트 문서 집합 UA-2에서의 분류 성능

UA-2에서의 분류 성능은 UA-1에 비해서 전반적으로 낮았다. Figure 11 와 Table 18 은 테스트 문서 집합 UA-2에 대해서 네 종류의 특징 그룹을 조합하여 실험한 결과이다. 특징 그룹 중 가장 높은 성능을 보여준

것은 BOW로 73.7%의 Accuracy를 기록하였으며, 다음으로는 언어적 특징 그룹이 69.2%의 Accuracy를 보였다. 두 특징 그룹의 차이점은 언어적 특징 그룹이 Results와 Methods 섹션에서 높은 성능을 보였으며, BOW는 Introduction에서 가장 높은 성능인 F-score 77.6%을 기록하였다. 반면, 구조적 특징은 가장 낮은 47.3%의Accuracy를 기록하였다.

두 개의 그룹을 조합한 경우에는, BOW가 사용된 조합이 전반적으로 높은 성능을 보였고, 언어적 특징은 구조적 특징과 조합되었을 때 77.0%의 accuracy로 높은 성능 향상을 보여주었다. 그러나 문법적 특징 그룹과 언어적 특징 그룹이 조합된 경우는 가장 낮은 69.4%의 Accuracy을 기록하였다. 세 개의 특징 그룹을 조합한 경우에는, BOW와 문법적 특징, 그리고 언어적 특징 그룹을 사용한 BGL이 가장 높은 성능인 77.9%의 Accuracy를 기록하였다. 반면, 언어적 특징 그룹을 사용하지 않은 BGS의 경우는 76.2%의 accuracy를 기록하여 가장 낮은 성능을 보여주었다. UA-2에서의 실험은 UA-1과는 달리 BOW와 언어적 특징 그룹이 같이 사용된 BL과 BGL에서 다른 조합에 비해 높은 성능을 보여주었다. BGL의 성능은 구조적 특징 그룹까지 포함된 BGSL (Accuracy = 77.4%)보다 0.5% 높은 성능을 기록하였다.

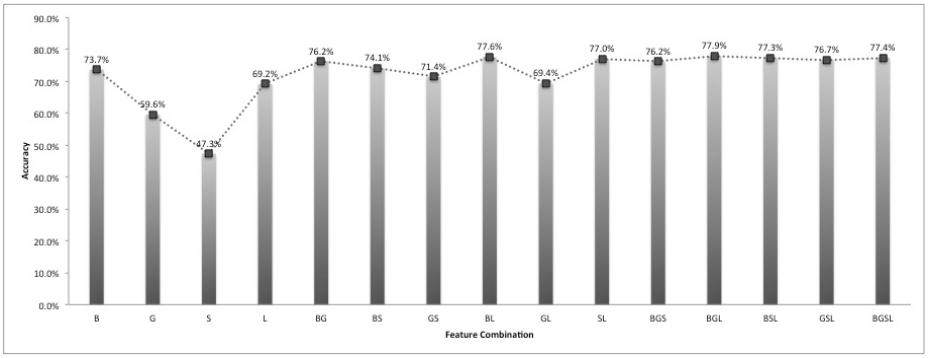

Figure 11 테스트 문서 집합 UA-2에서의 분류 성능 (Accuracy)

Table 18 테스트 문서 집합 UA-2에서의 분류 성능

Feature Combination	Accuracy	F-Score (Precision/Recall)			
		Introduction	Methods	Results	Discussion
B	73.7	77.6 (78.2/77.0)	58.8 (53.8/65.0)	77.5 (82.7/73.0)	68.6 (62.1/76.7)
G	59.6	59.9 (52.3/70.1)	47.1 (39.0/59.5)	67.2 (80.3/57.8)	49.7 (54.0/46.0)
S	47.3	75.9 (79.8/72.4)	30.4 (19.4/70.4)	47.7 (56.8/41.0)	7.8 (78.9/4.1)
L	69.2	65.2 (67.3/63.3)	65.1 (65.2/65.0)	78.4 (81.4/75.6)	54.6 (47.8/63.6)
BG	76.2	77.7 (80.3/75.2)	61.7 (57.8/66.1)	80.7 (82.8/78.6)	72.2 (67.1/78.1)
BS	74.1	82.1 (81.5/82.6)	52.5 (42.5/68.6)	77.0 (85.7/69.9)	72.6 (69.9/75.6)
GS	71.4	79.8 (75.5/84.6)	52.3 (39.7/76.6)	70.6 (90.1/58.6)	78.0 (72.0/85.2)
BL	77.6	79.4 (81.7/77.3)	65.7 (60.7/71.5)	81.9 (86.2/78.0)	72.2 (64.8/81.4)
GL	69.4	65.5 (67.6/63.4)	65.5 (62.4/69.0)	78.7 (83.1/74.6)	54.8 (47.8/64.4)
SL	77.0	83.1 (78.2/88.7)	63.8 (53.3/79.6)	76.9 (91.8/66.1)	77.7 (70.0/87.4)
BGS	76.2	81.7	59.2	79.0	73.8

		(79.7/83.7)	(49.4/73.7)	(86.4/72.8)	(72.8/74.8)
BGL	77.9	79.4 (82.9/76.1)	66.9 (61.0/74.1)	82.3 (85.8/79.1)	72.3 (65.6/80.5)
BSL	77.3	83.2 (82.0/84.4)	59.5 (50.4/72.6)	79.1 (87.3/72.3)	77.7 (72.9/83.3)
GSL	76.7	81.6 (75.2/89.3)	67.7 (58.1/81.0)	76.7 (91.8/65.8)	75.3 (68.0/84.4)
BGSL	77.4	81.1 (78.3/84.1)	61.2 (52.3/73.7)	80.1 (88.3/73.2)	77.2 (74.1/80.5)

언어적 특징의 Confusion matrix

Table 19 은 테스트 문서 집합 3개에 대해서 언어적 특징 그룹만을 사용했을 때의 Confusion matrix들이다. 언어적 특징 그룹은 모든 테스트 문서 집합에서 다음과 같은 특징을 보여주었다. 첫 번째, Introduction, Methods, 그리고 Results 섹션에서의 분류 성능이 Discussion보다 항상 좋았다. 그리고 UA-1의 Results섹션에서는 가장 좋은 분류 성능을 보여주었다. 두 번째, Introduction은 주로 Discussion으로 잘못 분류되었으며, 그 반대의 현상도 나타났다. 세 번째, Methods 섹션은 모든 테스트 문서 집합에서 Results 섹션으로 빈번하게 오 분류되었다.

또한, 언어적 특징이 오 분류된 문장에서 어떤 역할을 했는지 추출된 특징을 살펴본 결과, N-gram과 동사구의 가중치 순위가 불일치 하는 경우가 많음을 알 수 있었다. 이러한 불일치의 예로서는 다음과 같은 문장이 대표적이다.

To investigate whether providing videotelephone-based support was acceptable to these families, a 12-month

non-randomised acceptability trial was completed.

위 문장은 어노테이터가 Methods 섹션으로 분류된 것이었지만, 추출된 특징을 보면 n-gram “To investigate whether”와 동사구 “was completed”가 각각 Introduction과 Methods에서 가중치가 높아 서로 다른 순위 벡터를 제공하였다. 이렇게 두 섹션의 특징을 모두 포함하는 현상은 UA-1보다 UA-2에서 보다 많이 발견되어, UA-1을 구성하는 RCT 초록들이 문장 표현에서 UA-2보다 구조화된 초록과 유사함을 알 수 있었다.

Table 19 각각의 테스트 문서에 대한 언어적 특징의 Confusion matrix

(a) Structured Abstract Dataset

Actual Section	Predicted Section			
	Introduction	Methods	Results	Discussion
Introduction	77.10%	5.63%	5.51%	11.75%
Methods	6.34%	79.76%	12.73%	1.17%
Results	3.91%	8.05%	83.56%	4.49%
Discussion	20.28%	2.47%	22.93%	54.32%

(b) Unstructured Abstract Dataset-1

Actual Section	Predicted Section			
	Introduction	Methods	Results	Discussion
Introduction	69.31%	10.23%	6.47%	13.99%
Methods	5.31%	79.70%	14.04%	0.95%
Results	3.01%	6.78%	87.19%	3.01%
Discussion	15.66%	2.41%	15.26%	66.67%

(c) Unstructured Abstract Dataset-2

Actual Section	Predicted Section			
	Introduction	Methods	Results	Discussion
Introduction	63.29%	6.19%	12.84%	17.67%
Methods	10.95%	64.96%	22.26%	1.82%
Results	8.51%	4.26%	75.62%	11.61%
Discussion	21.37%	1.64%	13.42%	63.56%

6. 논의

본 연구는 의생명 분야의 언어적 특징을 이용하여 비구조화된 초록의 문장에 IMRAD중 하나의 섹션명을 태깅하는 것을 목적으로 하였다. 이를 위해 대규모의 구조화된 초록에서 IMRAD의 각 섹션을 대표할 수 있는 언어적 특징을 추출하여 이를 문장 분류에 사용하였고, 실험을 통해 언어적 특징 그룹이 다른 특징 그룹들과 같이 사용되었을 때 분류 성능을 개선시킬 수 있음을 보였다. 성능 면에서 보면, 언어적 특징 그룹을 사용하지 않는 조합보다1.8%에서 6.3%의 성능향상을 보여주었다. 즉, 본 연구에서 제안한 방법이 기존의 분류 방법을 개선시킴을 보였다. 또한 계산 비용 관점에서 보면, 제안한 방법에서는 15차원의 언어적 특징으로도 1,541차원의 BOW와 비슷한 성능을 낼 수 있었으며, 이러한 결과는 언어적 특징이 BOW를 대체할 수 있을 뿐 아니라 적은 계산 비용으로도 좋은 분류 성능을 내는 시스템 개발이 가능함을 보여주었다.

실험 결과는 또한 언어적 특징 그룹이 기존의 다른 특징 그룹들과 함께 사용되었을 때 눈에 띄는 성능향상이 있음을 보여주었다. 특히, 구조적 특징 그룹과 함께 사용되었을 때 가장 높은 성능 향상을 보여주었다. 언어적 특징 그룹 중에서는 N-gram이 가장 좋은 분류 성능을 보여주었는데, n의

값을 2~6의 범위로 사용했을 때가 가장 좋은 성능을 냈다. n의 범위를 7 이상으로 하는 것은 분류 성능에 큰 영향을 끼치지 않았다.

본 연구에서 언어적 특징 그룹을 이용한 최대 분류 성능은 SA에서 Accuracy 91.7%와 RCT 초록인 UA-1에서Accuracy 86.3%, 그리고 일반적인 비구조화된 초록 UA-2에서 Accuracy 77.9% 이었다. 이러한 기록은 대상 테스트 문서 집합의 특징을 잘 나타내는데, 비구조화된 초록은 반드시 IMRAD 형식에 따라 작성되지 않으므로 분류 성능이 구조화된 초록인 SA보다는 낮았지만, UA-1를 구성하는 RCT 초록의 경우는 IMRAD에 가까운 문장 구성으로 상대적으로 분류하기에 적합함을 알 수 있었다. UA-2의 경우는 섹션별 문장 분포로 볼때 연구 결과의 보고에 집중하는 경향이 잘 나타났으며, 이로 인해 문장 위치와 같은 구조적 특징이 좋은 분류 특징으로 사용될 수 없었다.

본 연구는 비구조화된 초록의 문장을 IMRAD중의 한 섹션명으로 자동적으로 태깅할 수 있는 기술을 제공한다는 점에서, 여러 가지 실제적인 시스템 구현의 첫걸음을 제공한다. 먼저 비구조화된 초록을 구조화된 초록으로 자동으로 변환하는 시스템이 가능하며, 이를 통해 연구자들이 논문을 더욱 정확하게 검색할 수 있을 것이다. 또한 본 연구에서 구축한 IMRAD 섹션의 각 언어적 특징들은 초록 작성을 돕는 시스템(Jeong, Nam et al. 2014)에서 중요한 역할을 할 수 있을 것이다.

Ⅳ. 의생명 초록 문장 자동 태깅 시스템

1. 시스템 소개

본 장에서는 2장과 3장에서 설명한 언어적 특징과 문장 분류 기술을 이용한 자동 태깅 시스템을 소개한다. 또한 IMRAD 초록의 각 섹션을 대표하는 언어적 특징을 빈도수를 기준으로 정리하여, 사용자가 각 섹션을 대표하는 특징을 조회할 수 있는 기능을 소개한다. 이러한 기능들은 연구자들의 편의를 위해 웹 서비스(http://abstract.bike.re.kr) 로 구축되었으며, 제공되는 메뉴를 통해 해당 기능을 사용할 수 있다.

2. 서비스 구성

2.1. INTRODUCTION

INTRODUCTION 메뉴는 RESEARCH OBJECTIVE와 DATA SET 메뉴로 구성된다. INTRODUCTION 메뉴에서는 연구 목표를 설명하고, DATASET 메뉴에서는 구조화된 초록의 언어적 특징을 추출하기 위해 사용한 코퍼스와 문장 분류 시스템을 평가하기 위해 구축한 3종류의 테스트 문서 집합에 대한 정보를 제공한다. 각각의 메뉴에 대한 화면은 Figure 12와 Figure 13과 같다.

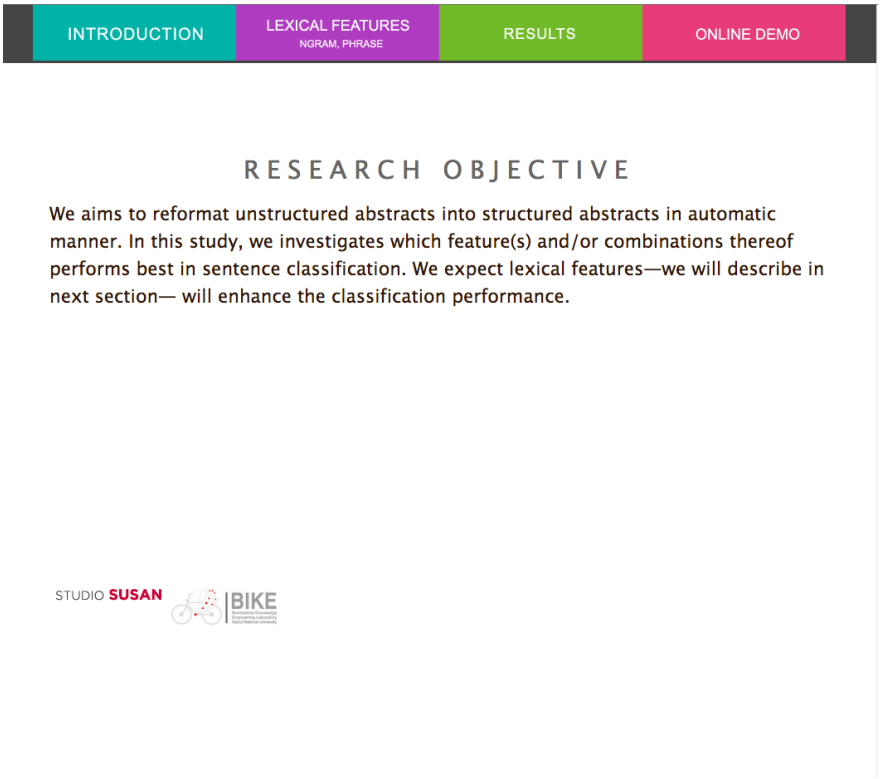

Figure 12 연구 목표 소개

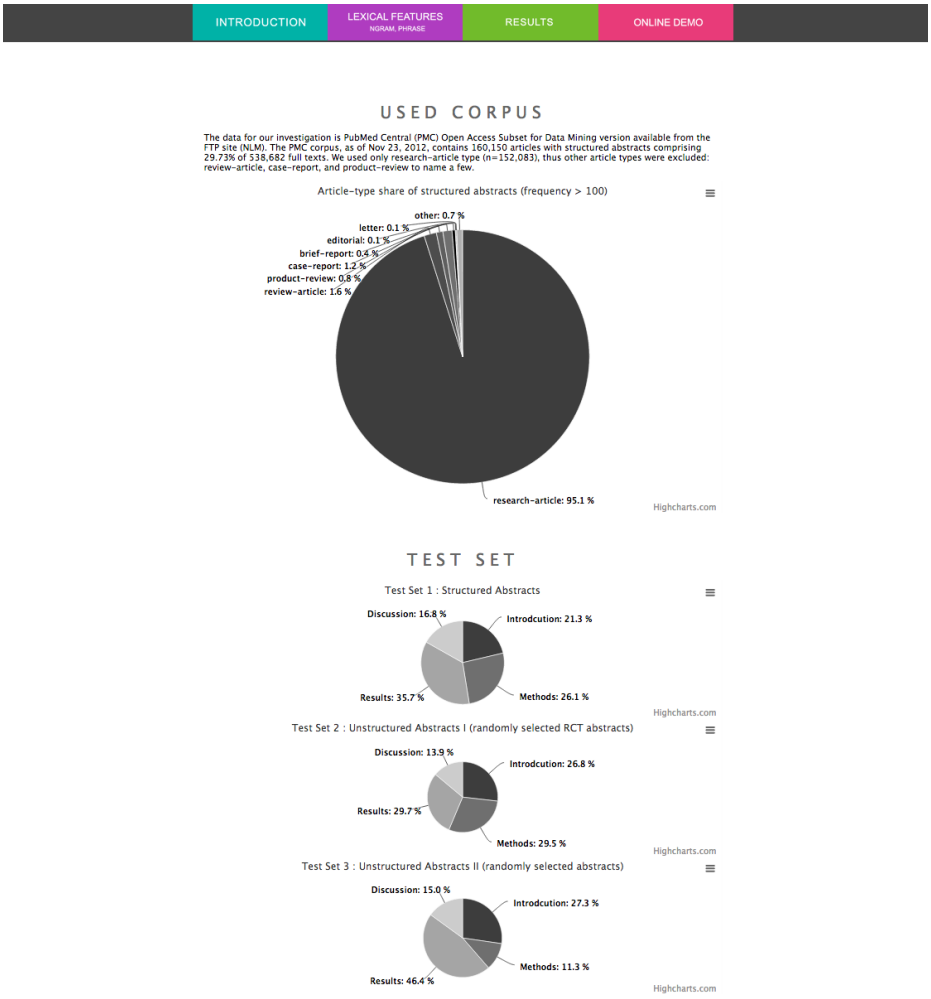

Figure 13 사용 코퍼스 및 테스트 문서집합

2.2 LEXICAL FEATURES

LEXICAL FEATURES 메뉴에서 사용자는 구조화된 초록의 각 IMRAD 섹션에서 추출한 N-gram과 명사/동사구의 빈도수를 조회할 수 있다. 사용한 구조화된 초록은 코퍼스에서 포함된 research-article 타입

의 구조화된 초록 152,083개이며, 초록에 사용된 다양한 섹션명은 2장에서 언급한 정규화 및 맵핑 과정을 통해 최종적으로 IMRAD섹션으로 맵핑된 것이다. 먼저, 서브 메뉴 "KEY NGRAM LIST" 에서는 각각의 IMRAD섹션에서 추출한 N-gram과 사용된 빈도수를 섹션별로 조회할 수 있는 기능을 제공한다. 사용자는 n의 크기를 지정하여 검색할 수 있어, 각 섹션의 대표적인 패턴 및 문형을 검색할 수 있다.

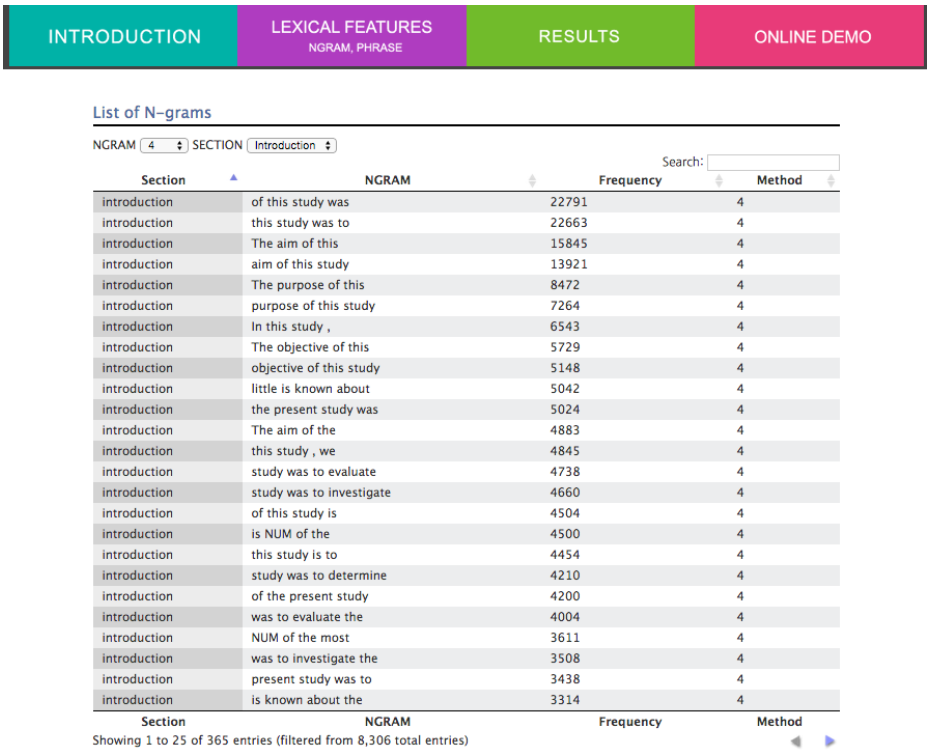

Figure 14 섹션별 핵심 N-gram 조회

서브메뉴 "KEY PHRASE LIST" 에서는 섹션별 핵심 명사구와 동사구를 조회할 수 있다. 사용자는 "PHRASE TYPE"에서 조회 대상을 "ALL",

"Noun Phrase", 그리고 "Verb Phrase"에서 선택할 수 있으며, N-gram과 마찬가지로 조회 대상 섹션을 선택할 수 있다. 검색 결과는 빈도수 기준으로 정렬되어 제공된다.

Figure 15 섹션별 핵심 명사/동사구 조회

2.3 RESULTS

RESULTS 메뉴에서는 언어적 특징과 기존 연구에서 사용된 특징 그룹들을 조합하여 실험한 3장의 결과를 그래프로 제공하고 있다. 그래프는 4개가 제공되며, 각각은 언어적 특징 각각의 문장 분류 성능, 구조화된 초

록으로 이루어진 테스트 문서 집합(SA)에서의 분류 성능, 비구조화된 RCT 초록으로 구성된 테스트 문서 집합(UA-1)에서의 분류 성능, 그리고 일반적인 비구조화된 초록으로 구성된 테스트 문서 집합(UA-2)에서의 분류 성능을 표시한 것이다.

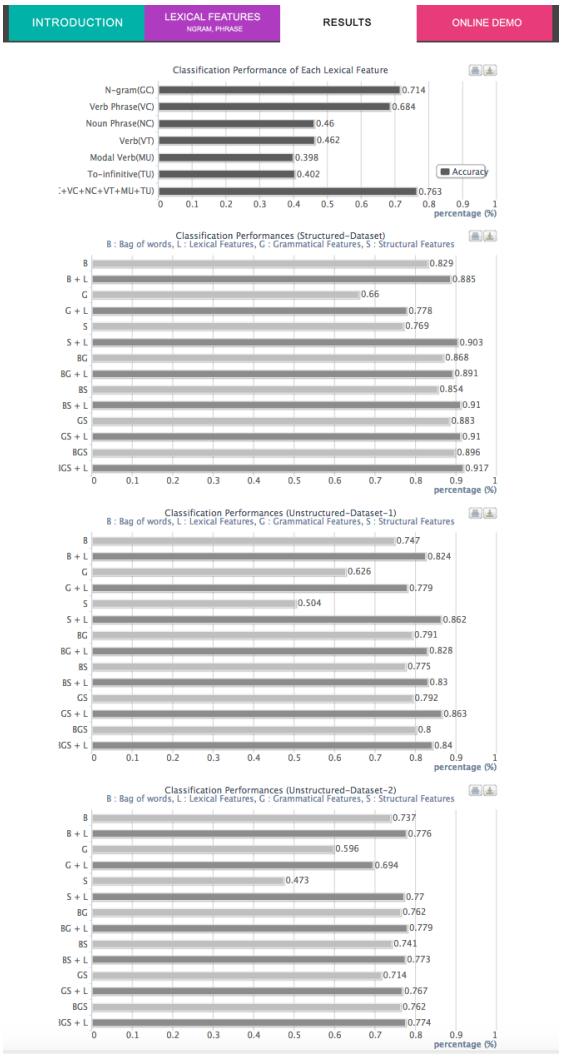

Figure 16 문장 분류 성능 그래프

2.4 ONLINE DEMO

ONLINE DEMO 메뉴에서는 3장에서 설명한 연구 결과를 이용하여 개발한 자동 태깅 시스템을 소개한다. 초기 화면은 PubMed에서 가져온 임의의 비구조화된 초록을 입력으로 사용하였으며, 초록 전체를 자동 태깅하도록 설정되었다. 사용자는 "Classify" 버튼을 클릭하므로 각각의 문장을 자동 태깅할 수 있다.

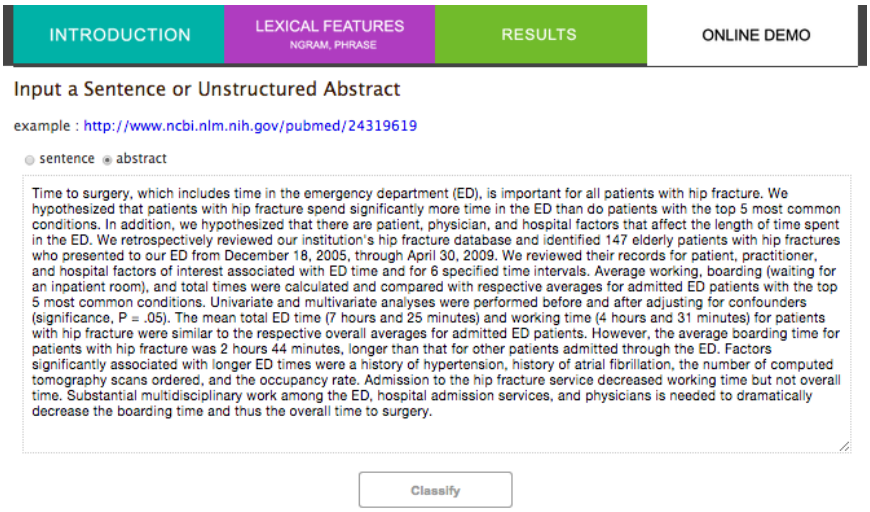

Figure 17 자동 태깅 시스템 초기화면 (초록)

Figure 17에서 사용한 초록은 PubMed ID가 24319619에 해당하는 것으로, 구조화된 초록의 형식을 따르지 않는 초록이다. 본 시스템은 각 문장을 분류하기 위하여 3장에서 학습시킨 모델 중에서, 언어적 특징 그룹, 문법적 특징 그룹, 그리고 구조적 특징 그룹을 사용한 것을 사용하였다.

문장의 자동 태깅을 위해서는 입력 초록을 문장으로 구분한 후에, 각 문장을 대상으로 분류 특징을 추출하고 이를 이용하여 문장 분류를 위한 특징 벡터를 구축하였다. 문장에서 추출된 특징 벡터는 자동 태깅 시스템의 입력으로 사용되며, 그 결과는 IMRAD섹션중 하나의 값이다. Figure 18 은 자동 태깅 시스템의 결과인데, 각 문장은 IMRAD 섹션 중 하나로 태깅된 것을 볼 수 있다.

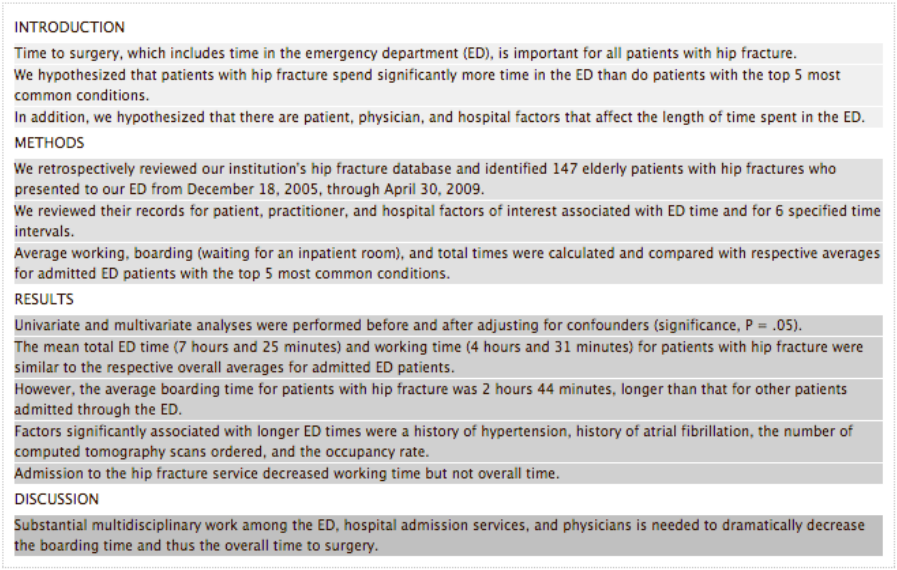

Figure 18 자동 태깅 결과

자동 태깅 시스템의 입력은 초록뿐 만 아니라 문장이 될 수 도 있다. 사용자는 메뉴에서 "sentence" 를 선택하여 문장을 자동 태깅할 수 있다.

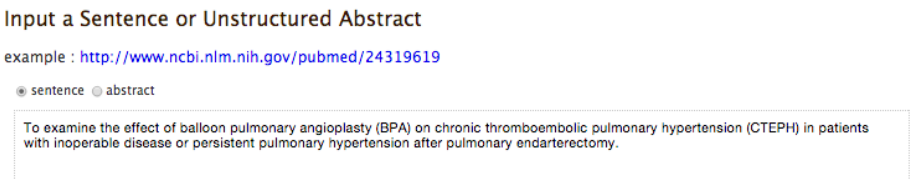

Figure 19 문장 태깅 선택

Figure 20 은 문장을 입력으로 받아 자동 태깅한 결과이며, 사용된 모델은 3장에서 문법적 특징 그룹과 언어적 특징 그룹으로 학습시킨 모델을 사용하였다.

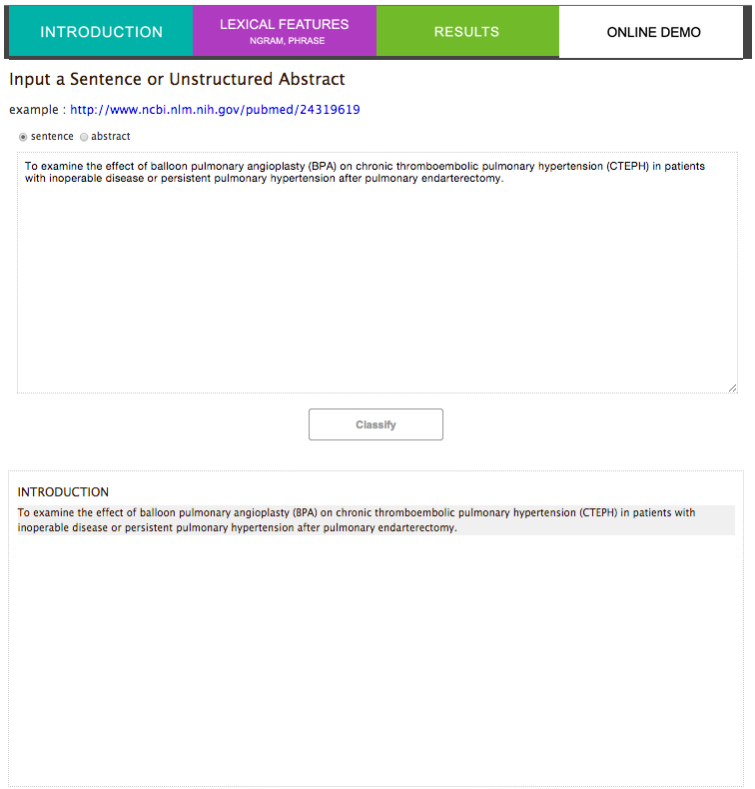

Figure 20 문장 태깅 결과

3. Use Cases

비구조화된 초록을 대상으로 한 자동 태깅 시스템은 여러 분야에서 검색 정확률을 높이거나 효과적인 정보 추출이 가능하게 하는 기반 기술로 활용될 수 있다. 본 자동 태깅 시스템의 활용 예를 보면 다음과 같은 것들이 있을 수 있다.

첫 번째, 논문 검색을 더욱 정확하게 하는데 사용될 수 있다. 현재 의생명 분야에서 논문 검색을 위해 가장 많이 사용되는 PubMed (http://www.ncbi.nlm.nih.gov/pubmed)의 경우, 검색 결과는 Figure 21와 같다. 검색 결과는 사용자가 입력한 키워드와 관련된 논문들로서 사용자는 검색된 논문이 관련된 것임을 확인하기 위해서는, 해당 논문의 링크를 클릭한 후 초록을 읽어보아야 한다. 이때 초록이 비구조화된 초록의 경우라면 초록 전체를 읽는 과정이 포함된다.

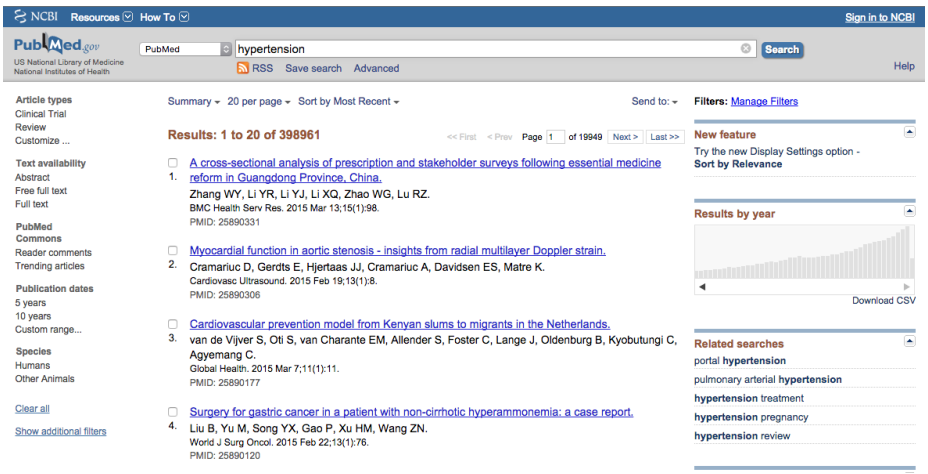

Figure 21 PubMed 검색 결과 예

만약 자동 태깅 시스템을 적용한다면, 각각의 논문의 초록은 구조화된 초록으로 변환된 후 색인될 수 있을 것이다. 그래서, 사용자는 논문의 특정 섹션을 대상으로 검색할 수 있게 되어 더욱 정확한 검색이 가능하며, 기존의 논문 확인 과정을 단순화시킬 수 있을 것이다. 예를 들어, 연구 목적이 포함된 Introduction을 대상으로 논문을 검색하거나, 특정 Methods 섹션을 대상으로 특정 연구 방법을 사용한 논문을 더욱 정확하게 검색할 수 있다. 검색 결과를 보여주는 화면 역시, 사용자가 Introduction섹션을 대상으로 검색한 경우에는 논문의 제목과 Introduction을 같이 보여주어 논문 선택이 효율적이게 된다.

두 번째, 초록에서 더욱 정확한 정보 추출이 가능하게 한다. 메타 분석과 같이 논문의 동향을 분석하는 분야에서는 논문에 사용된 키워드나 부착된 MESH용어가 중요한 정보로 사용된다. 이때, 본 연구의 자동 태깅 시스템을 사용한다면, 키워드가 어떤 섹션에서 나왔는지를 알 수 있기 때문에 더욱 다양한 분석이 가능할 수 있다. 예를 들어, 초록의 Methods섹션에서 사용된 키워드의 동향을 분석한다거나, Introduction에 포함된 키워드의 동향 분석은 기존의 방법과는 다른 관점의 분석을 가능하게 할 것이다.

Ⅴ. 구조적 특징을 이용한 임상 서식의 태깅

1. 연구 배경

EMR의 도입에 따라 더욱 더 많은 임상 문서들이 전자화된 형태로 사용되고 있는 가운데(Vawdrey 2008), 임상 문서들의 구조를 정의한 임상서식의 중요성 역시 증가하고 있다. 다양한 임상 문서의 사용으로 인한 임상 서식(CDT, Clinical Document Template)의 증가는, 서식의 재사용성과 접근성의 개선을 요구하고 있다. 임상 서식과 관련된 기존의 연구들 중에 대표적인 것들은 주로, 임상 서식을 구성하는 정보를 어떻게 모델링 할 것인가에 대한 것인데, 대표적인 결과물로는 Detail Clinical Model (DCM)(Coyle, Mori et al. 2003, Goossen, Goossen-Baremans et al. 2010), Clinical Element Model (CEM)(Coyle, Heras et al. 2008), Clinical Contents Model[4](CCM), 그리고 openEHR Archetypes(Beale 2003, Chen, Klein et al. 2009, Garde, Chen et al. 2009)이 있다. 이 정보모델들은 임상에서 사용되는 개념들과 이들간의 관계를 표준화된 형식으로 구조화하여 재사용성을 증가시키는 도구로서 사용되어왔다. 예를 들어 CEM의 경우에는, 임상 개념과 관련된 값들을 위해서 Abstract Instance Model을 사용하며, 개념들간의 관계는 Semantic Link를 이용하여 정의하였다. CCM의 경우는, 임상 개념을 표현하기 위해 Entity, Qualifier/Modifier, 그리고 Value로 구성된 EQV 구조를 정의하였으며,

[4] http://ccm.hins.or.kr/main.php

다양한 언어로 개념명을 표시하고 이를 SNOMED-CT [5](Systematized Nomenclature of Medicine – Clinical Terms)와 같은 표준화된 용어와 맵핑할 수 기능을 제공하였다. openEHR은 Archetype이라는 재사용 가능한 구조를 사용하는데, ADL(Archetype Definition Language)을 이용하여 새로운 Archetype을 정의할 수 있도록 하였다.

그러나, 이러한 정보 모델은 임상 서식에 사용되는 정보들을 모델링 하는데 초점을 맞추고 있으며, 서식을 작성하는 행동에 내재된 지식들은 모델링의 대상에 포함시키지 않았다. 임상 서식을 작성할 때는 서식의 목적에 부합하는 서식 항목을 선택하고 배치하는 행동이 필수적으로 포함된다. 즉, 임상 서식은 해당 분야의 전문 지식이 사용된 지적 결과물인 것이다. 그리고, 이러한 서식 작성은 위에서 언급한 서식 항목의 선택과 배치라는 두 가지 중요한 행동을 포함한다. 이러한 행동의 결과는 여러 가지 새로운 정보 모델의 필요성을 요구한다. 예를 들어 "Hypertension"과 "Tuberculosis"은 임상 개념 레벨에서는 서로 관련되어 있다고 볼 만한 여지가 없지만, 특정 과의 (예를 들어, 소화기 내과) "Admission Note"란 서식에서 "Past Illness" 의 세부 항목으로 같이 사용되었다면 "Hypertension"과 "Tuberculosis"은 "Past Illness"란 맥락에서 중요한 관계를 맺고 있음을 알 수 있다. 본 연구에서는 바로 이러한 지식을 서식에서 추출하고, 이를 재사용 가능한 형식으로 태깅하기 위한 것이다. 본 연구에서는 이를 위해 지식 표현을 위해 사용되는 온톨로지 기술을 사용하였다.

---

[5] http://www.ihtsdo.org/snomed-ct

2. 연구 목표

본 연구에서는 임상 서식의 작성 행위에 내재된 지식을 태깅하는데 사용할 새로운 지식 모델을 제안하고, 이를 온톨로지 기술을 이용하여 표현하였다. 개발된 온톨로지는 임상 서식에서 추출된 지식을 태깅하는데 사용되어, 새로운 서식을 작성하거나 서식을 검색할 때에 효과적으로 재사용될 수 있도록 한다(Kim, Ha et al. 2005).

3. 임상 서식의 태깅을 위한 지식 모델

3.1. 온톨로지

본 연구에서는 임상 서식의 작성 과정에 내재된 지식을 태깅하기 위한 지식 모델과 이를 재사용 가능한 형식으로 표현하기 위해 온톨로지를 사용하였다. 의생명 분야에서 온톨로지는 지식 표현의 수단으로서 광범위하게 사용되고 있다. 위에서 언급한 archetype이나 CEM(Tao, Parker et al. 2011)도 정보 모델을 온톨로지를 이용한 예중의 하나이다. 또한 ADL을 W3C의 OWL[6]로 변환한 (Martínez-Costa, Menárguez-Tortosa et al. 2009, Lezcano, Sicilia et al. 2011)의 연구에서는 온톨로지 기술을 사용하여 archetype의 비교, 선택, 분류, 그리고 일관성 검사를 효과적으로 할 수 있도록 하고, 나아가 Archetype Object Model (AOM) 의 객체간 관계를

[6] http://www.w3.org/TR/owl2-overview/

통해 추론이 가능하도록 하였다.

본 연구에서도 임상 서식에서 추출한 지식을 온톨로지로 표현하기 위하여 W3C의 웹 온톨로지 언어인 OWL을 사용하였다. OWL은 개념을 표현하기 위한 클래스와 개념의 구체적인 예에 해당하는 인스턴스, 그리고 개념의 특성 및 개념간 관계를 표현하기 위한 프로퍼티를 제공한다. 또한 W3C에서는 OWL을 이용한 온톨로지를 공개하고 있는데, 출처(Provenance)를 표현하기 위한 PRO-V 온톨로지[7]가 그 중의 하나이다. 본 연구에서는 임상 서식의 출처와 변경 사항을 표현하기 위하여 PRO-V 온톨로지를 사용하였다. PRO-V 온톨로지는 출처와 관련된 prov:wasGeneratedBy나 prov:WasDerivedFrom와 같은 프로퍼티를 제공할 뿐 아니라, 특정 활동의 시작과 끝을 표현하기 위한 prov:startedAtTime과 prov:endedAtTime 프로퍼티를 제공하고 있다. 이외에도 PRO-V 온톨로지는 임상 서식을 작성하는 과정에서 Copy & Paste 로 발생하는 서식간의 관계를 표현할 수 있는 다양한 프로퍼티들을 제공한다.

3.2. 개념 모델

Figure 22은 본 연구에서 추출하고자 하는 지식을 설명하기 위해 임상 서식 "Admission Note"의 한 부분을 뽑아온 것이다. 그림에서 서식 작성자는 "Admission Note" 서식에서 필요한 정보를 채우기 위해, 서식 항목

[7] http://www.w3.org/TR/prov-o/

"Past Medical History", "Social History", 그리고 "Family History"를 사용하였으며, 다시 "Past Medical History" 아래에 "DM", "Hypertension", "Hepatitis", 그리고 "Drug Allergy"를 두었다.

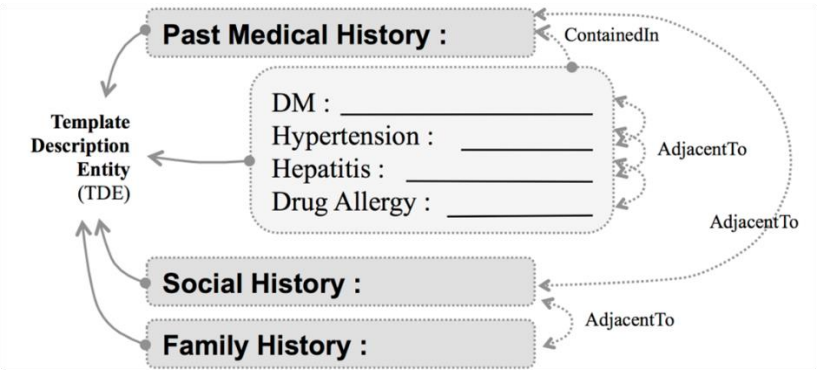

Figure 22 임상서식의 한 부분

서식 항목과 서식 항목 사이에는 배치를 통해 관계가 만들어질 수 있는데, 위의 예에서는 "Past Medical History"와 하위 4개의 항목과는 포함관계가, "DM"과 "Hypertension"은 동일한 수준에서 사용되어 인접 관계가 생성될 수 있다. 이러한 관계는 "Admission Note"라는 맥락에서 중요한 지식으로 사용될 수 있는데, 위에서 대표적인 예로는 "Past Medical History"와 "DM", "Hypertension", "Hepatitis", 그리고 "Drug Allergy"이 포함 관계란 점이다. 이 관계가 중요한 것은, 환자의 과거력중 위의 4가지를 사용한 것이 서식의 목적과 서식을 사용하는 특정 과(Department)와 연관 지었을 때, 중요한 지식으로 활용될 수 있기 때문이다. 본 연구에서 임상 서식에서 추출하려는 것이 바로 이러한 것들이며, 이러한 것들은 서식의 구조를 태깅하면서 가능하다.

개념 모델은 임상 서식에서 추출하려는 지식을 온톨로지로 표현하기 전에 주요 개념들을 도출하고 개념들간의 관계를 도식화 한 것으로, 본 연구에서는 Figure 23과 같이 정의하였다.

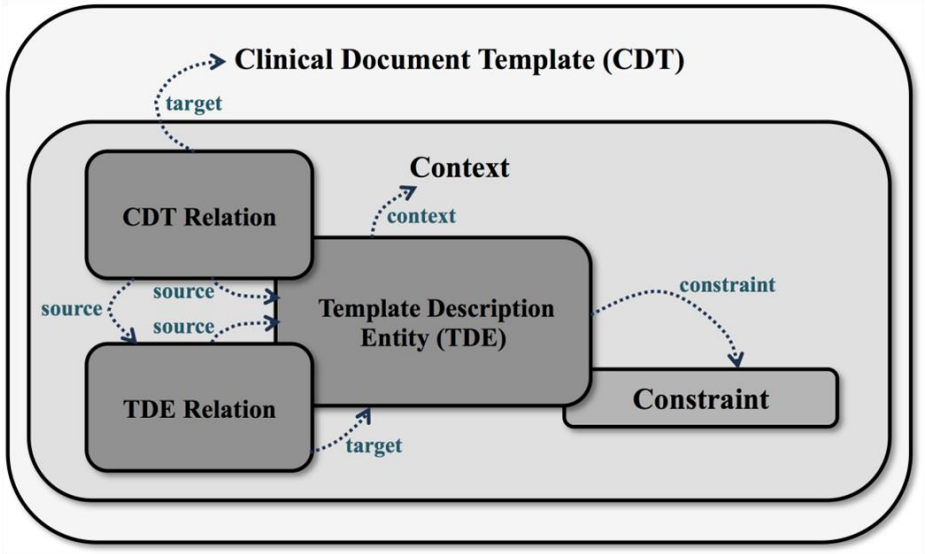

Figure 23 개념 모델

개념 모델에서는 임상 서식 작성 과정에서 가장 핵심이 되는 서식 항목의 선택과 서식 항목의 배치를 표현하기 위해, 서식 항목에 해당하는 Template description entity(TDE) 와 서식 항목의 배치를 통해 발생하는 관계를 나타내는 TDE Relation, 그리고 임상 서식을 나타내는 Clinical document template (CDT)을 정의하였다. 또한 TDE가 서식에 사용될 때 가질 수 있는 Constraint와 임상 서식 CDT와 서식 항목 TDE간의 관계를 표현하기 위한 CDT Relation을 정의하였다.

개념 모델에서 정의한 주요 개념들은 위에서 언급한 DCM, CEM, 그리

고 CCM에서 정의한 것과 대상은 비슷하나, 개념 모델이 표현하고자 하는 지식과는 차이점이 있다. 이러한 개념 모델의 특징은 Figure 24을 통해서 설명될 수 있다.

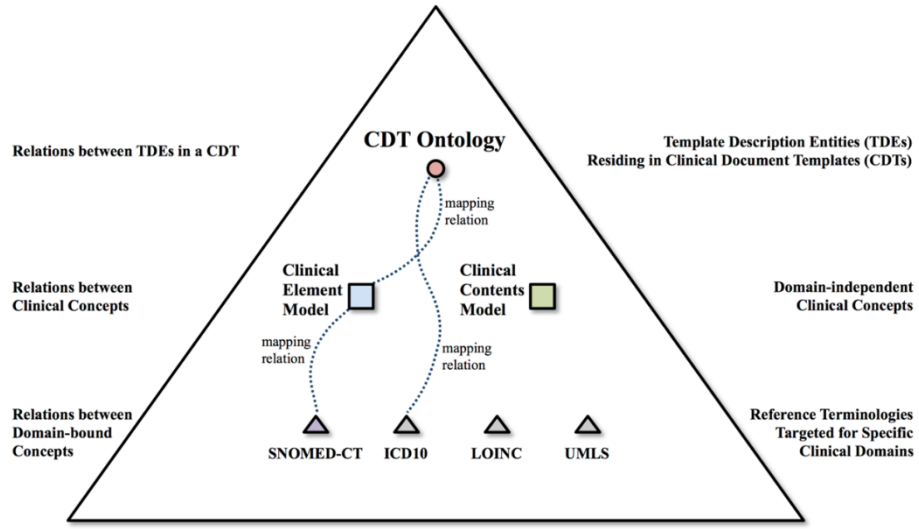

Figure 24 개념모델의 위상

Figure 24의 가장 아래 계층은 참조 용어 체계에 해당하는데, 이 계층은 일반적으로 질병이나 치료와 같은 특정 도메인을 위한 것이다. SNOMED-CT나 질병분류를 위한 ICD-10[8](International Classification of Diseases)등과 같은 것들이 이 계층에 해당한다. 그 다음 계층으로 올 수 있는 것들이 CEM이나 CCM 같은 정보 모델이다. 이 계층에서는 도메인 독립적인 임상 개념들은 아래 계층의 표준 용어 체계를 사용하여 정의하며,

[8] http://apps.who.int/classifications/icd/en/

서식 작성을 위한 재사용 가능한 정보 블럭으로 사용된다. 마지막으로 본 연구에서 제안하는 모델은 가장 상위에 올 수 있는 것으로, 임상 서식이라는 관점에서 기존의 정보 모델과 용어체계를 사용한다는 점에서 차이가 있다. 즉, 서식을 구성하는 TDE는 기존 모델의 임상 개념이거나, 참조 용어를 사용하여 새롭게 정의한 것일 수도, 아니면 로컬 용어로 정의하고 표준 용어를 참조할 수는 유연함을 가진다. 따라서, 본 연구에서 제안하는 모델은 기존의 임상 개념 모델을 대체하는 것이 아닌, 기존의 정보 모델을 그대로 사용하면서 서식에 포함된 지식을 표현한다는 장점을 가진다.

3.3. CDT 온톨로지

CDT 온톨로지는 위에서 정의한 개념 모델을 OWL 문법으로 표현한 것으로, 개념은 클래스로 관계는 프로퍼티를 사용하여 정의되었다. 클래스는 11개의 클래스와 4개의 오브젝트 프로퍼티, 그리고 2개의 데이터 타입 프로퍼티로 구성되어 있다. 또한 CDT 온톨로지는 외부 온톨로지와의 연계를 위해 Dublin Core Metadata Initiative (DCMI)의 메타데이터 용어[9]와 FOAF (Friend of a Friend) 용어집합[10]을 사용하였으며, 이를 통해 임상서식 CDT와 서식 항목 TDE 에 출처 정보 및 메타 정보를 기술하였다. 또한 CDT온톨로지는 W3C에서 제공하는 DCMI의 메타데이터 용어와 W3C의

---

[9] http://dublincore.org/documents/dcmi-terms/

[10] http://www.foaf-project.org/

PRO-V 온톨로지(Provenance Ontology)간의 맵핑정보[11]를 이용하여. PRO-V 온톨로지에서 정의한 출처 정보로의 변환이 자연스럽게 일어날 수 있도록 하였다. Table 20은 CDT 온톨로지에서 정의된 클래스와 프로퍼티에 대한 설명이다.

Table 20 CDT 온톨로지의 클래스와 프로퍼티

구분	이름	설명
클래스	ClinicalDocumentTemplate	EMR 시스템에서 사용하는 임상 서식을 나타내는 클래스
클래스	TemplateComponent	임상 서식에 포함된 개념들의 상위 클래스로서, 서식 항목을 나타내는 TemplateDescriptionEntity 클래스와 서식 항목간 관계를 나타내는 TDERelation의 부모 클래스
클래스	TemplateDescriptionEntity	임상 서식을 구성하는 서식 항목을 위한 클래스
클래스	Relation	임상 서식간 관계나 서식 항목간의 관계를 표현하기 위한 클래스로, 모든 관계의 부모 클래스
클래스	CDTRelation	임상 서식간 관계를 위한 클래스
클래스	TDERelation	서식 항목 TDE간 관계를 위한 클래스
클래스	UsedAtRelation	서식 항목 TDE와 임상 서식 CDT 간의 관계로, TDE가 CDT에 사용되었음을 표현하기 위한 클래스
클래스	ContainedInRelation	서식 항목 TDE간의 관계 중 포함

11

		관계를 위한 클래스
클래스	AdjacentToRelation	서식 항목 TDE간의 관계 중 인접 관계를 위한 클래스
클래스	Context	서식 항목이 사용된 문맥을 표현하기 위한 클래스
클래스	Constraint	서식 항목이 가지는 제약조건을 표현하기 위한 클래스
오브젝트 프로퍼티	source	Relation 클래스를 도메인으로 갖는 오브젝트 프로퍼티로, 관계의 시작이 되는 인스턴스에 대한 URI를 값으로 가짐
오브젝트 프로퍼티	target	Relation 클래스를 도메인으로 갖는 오브젝트 프로퍼티로, 관계의 종점이 되는 인스턴스에 대한 URI를 값으로 가짐
오브젝트 프로퍼티	context	TemplateDescriptionEntity 클래스를 도메인으로 하고 값(Range)으로는 Context 클래스의 인스턴스URI 를 가짐
오브젝트 프로퍼티	constraint	TemplateDescriptionEntity 클래스를 도메인으로 하고 값(Range)으로는 Constraint 클래스의 인스턴스URI 를 가짐
데이터타입 프로퍼티	department	임상서식이 사용된 과를 위한 프로퍼티
데이터타입 프로퍼티	filename	임상 서식 파일 명을 위한 프로퍼티

Figure 25은 CDT 온톨로지를 시각적으로 표현한 것이다. 그림에서 타원은 클래스를 나타내고, 타원과 타원을 연결하는 실선은 CDT온톨로지에서 정의한 프로퍼티이다. 또한 점선으로 표현된 것은 외부 온톨로지에서 정의한 것을 사용한 경우를 나타낸다.

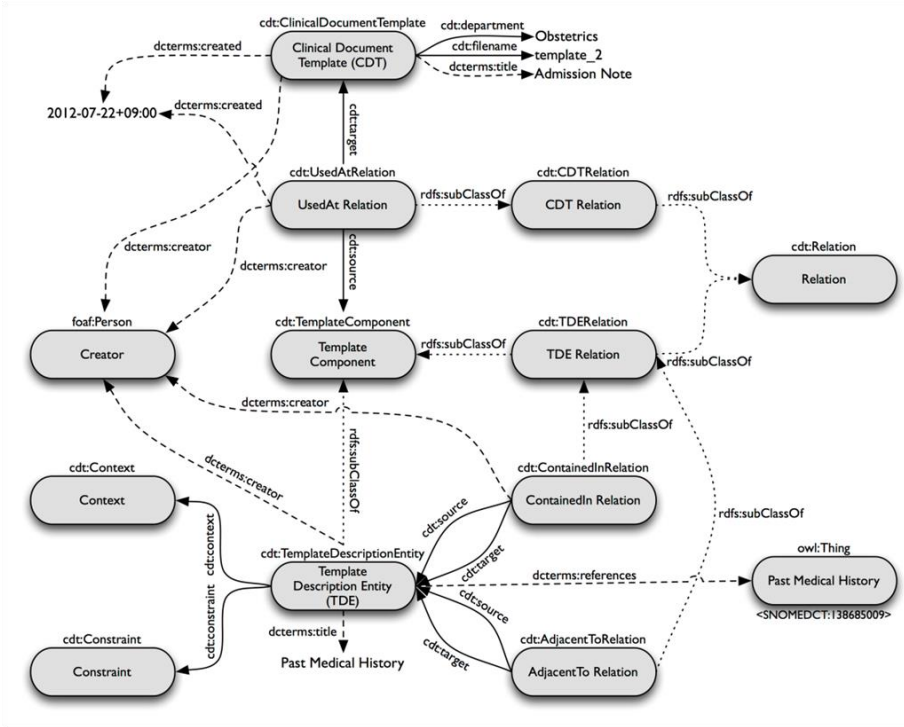

Figure 25 CDT 온톨로지

CDT 온톨로지는 다른 연구자들이 사용할 수 있도록 DERI에서 개발한 용어 출판 시스템인 Neologism [12] 을 이용하여 http://vocab.bike.re.kr/cdt# 에 공개하였다. 이 곳에서는 CDT온톨로지에서 정의한 클래스와 프로퍼티의 설명과 각각의 URI를 제공하여, 외부 온톨로지에서도 CDT 온톨로지에서 정의한 것들을 사용할 수 있도록 하였다. Figure 26 은 공개한 웹 사이트 화면이다.

[12] http://neologism.deri.ie

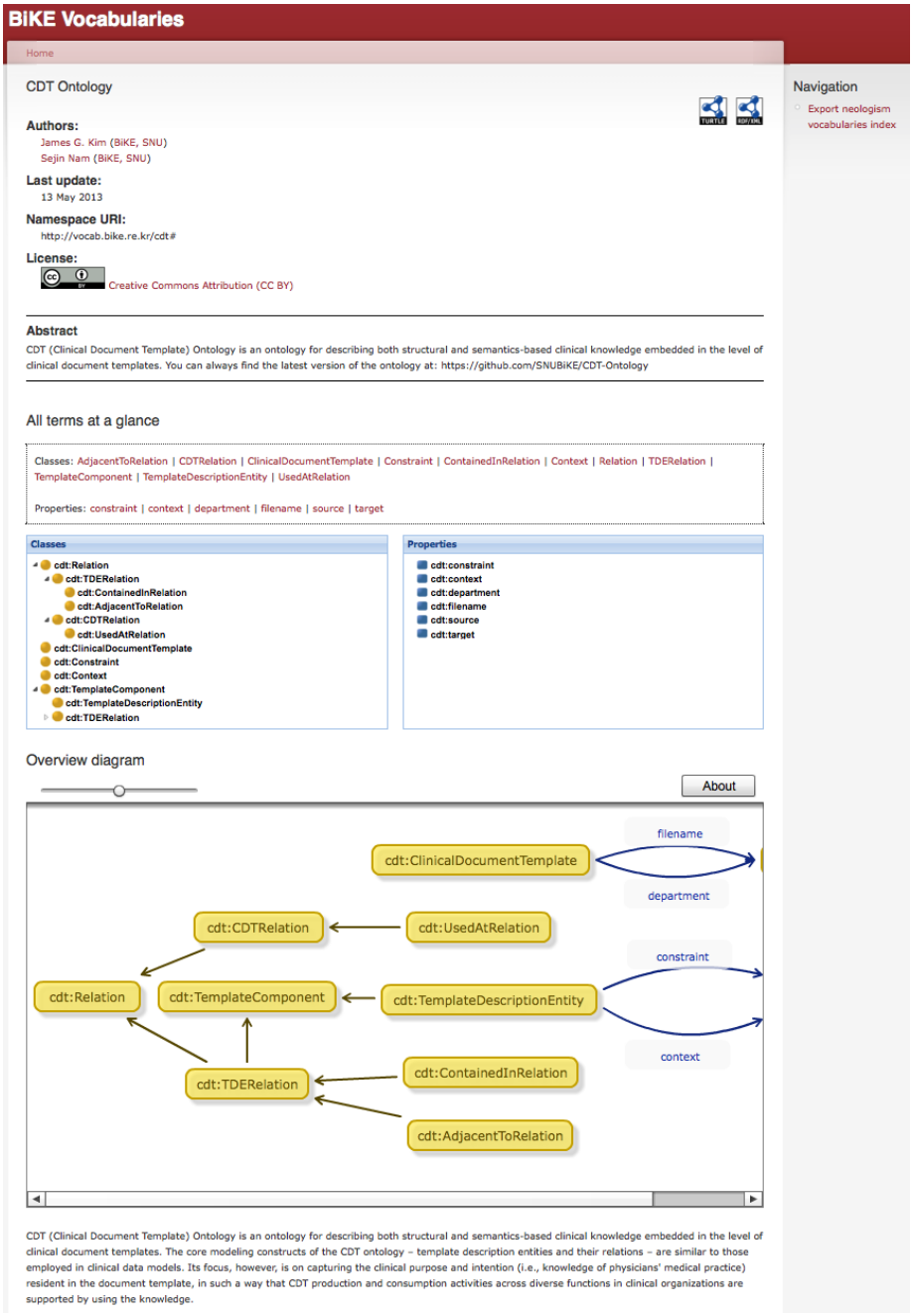

Figure 26 CDT 온톨로지 웹 사이트 화면

4. CDT 온톨로지를 이용한 임상서식 태깅

CDT 온톨로지를 이용한 임상 서식의 유용성을 테스트하기 위해, 본 연구에서는 의료기관에서 사용하는 임상 서식을 사용하여 태깅작업을 수행하였다. 사용된 서식은 500베드 규모의 병원에서 사용하는 것으로 입원, 퇴원, 간호, 수술, 치료 후 경과기록과 관련된 서식들 중에서 가장 많이 사용되는 35가지를 사용하였으며, 3명의 전문가가 참여하여 다음과 같은 규칙을 적용하여 임상 서식을 태깅하였다. 태깅된 결과물은 CDT 온톨로지의 인스턴스로 표현된 후, 트리플 저장소에 저장하였다.

- CDT (Clinical Document Template): 임상 서식당 CDT 클래스의 인스턴스를 만들고, 임상 서식의 이름과 파일 명을 데이터타입 프로퍼티를 생성한다.
- TDE (Template Description Entity): 임상 서식에 포함된 모든 키/값 쌍으로 표현된 서식 항목과 포함관계가 있는 서식항목을 추출하여 TDE 클래스의 인스턴스를 만든다. 예를 들어, 서식항목 "General Appearance"가 "Mental Status"와 "Looking Appearance"를 포함하는 경우, 각각에 대해서 TDE 인스턴스를 만들고 이들 간의 관계를 설정한다.
- Context: TDE가 사용된 서식의 타입 값을 사용한다. 본 연구에서는 "입원", "퇴원", "경과기록"등과 같은 값 중에 하나를 사용하였다.
- Constraint: TDE가 가질 수 있는 값의 범위나 TDE간에 존재하는 제약조건을 기록한다. 예를 들어, 임상 서식, 퇴원요약지에서 TDE "퇴원사유"가 사용되었고, 선택할 값으로는 "demise", "transfer", "discharge to

home", 그리고 "voluntary discharge" 가 있다면, "퇴원 사유"가 가질 수 있는 값의 범위는 위 4가지 중에 하나이며, 이를 Constraint로 표현한다. 그리고, 퇴원 사유가 "transfer"로 선택된 경우에만 TDE "hospital name"에 값을 입력하여야 한다면, 이와 같은 제약조건을 기록한다.

- CDT Relation: 임상서식 CDT와 TDE와 TDE 관계와 같은 서식 구성요소들간의 관계 UsedAt Relation을 기록한다. 임상 서식에 사용된 모든 TDE와 TDE간의 관계는 임상서식과 UsedAt Relation관계로 표현한다.
- TDE Relation: 임상서식에 나타난 TDE간의 관계 ContainedIn Relation과 AdjacentTo Relation을 기록한다. ContainedIn Relation은 특정 TDE가 하나 이상의 TDE를 포함하는 경우에 사용하며, AdjacentTo Relation 은 TDE가 서로 인접 관계일 때 사용한다.

이러한 태깅 규칙을 통해 35개의 임상서식에서 총 967개의 TDE를 추출하였는데, 이는 평균적으로 하나의 임상 서식에서 78개의 임상 정보를 얻을 수 있음을 알 수 있다. 이외에도 임상서식에서 추출한 인스턴스와 프로퍼티의 개수를 포함한 통계치는 Table 21 에 정리하였다.

Table 21 임상서식에서 추출한 인스턴스 구성

Category	Number of Instances
number of TDE instances	967
number of Contain Relations	1,412
number of Adjacent Relations	3,142
number of Used At Relations	7,283
number of TDE instances have “contain” relations	162
number of TDE instances have “adjacent” relations	932

number of TDE instances have both relations	150
average TDE instances of every templates	78.0
average contain relations of every TDE instance	8.7
average adjacent relations of every TDE instance	6.7

임상 서식에서 태깅을 통해 추출한 CDT 온톨로지 인스턴스들과 이들간의 관계는 Figure 27과 같이 시각적으로 표현될 수 있다.

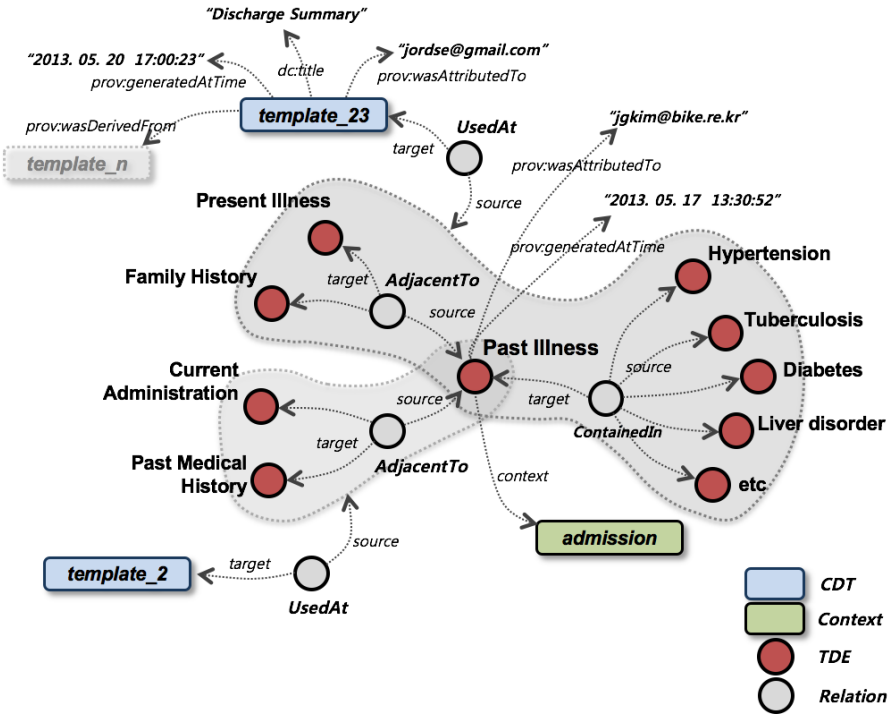

Figure 27 Past Illness와 다른 TDE 및 서식간의 관계

그림에서 TDE "Past Illness"는 "Family History", "Present Illness", "Current Administration", 그리고 "Past Medical History"등과 같은 TDE

들과 "template_1"과 "template_23"에서 "AdjacentTo Relation"으로 연결되어 있다. 또한 임상 서식 "template_23"은 2013년 5월 20일에 "Discharge Summary"란 이름으로 작성되었음을 알 수 있다. 이러한 관계는 "Past Medical History"란 TDE가 어떤 임상 서식에서 주로 사용되었는지를 알 수 있는 정보이며, 또한 어떤 TDE과 주로 사용되고 있는지를 알 수 있다. 또한 서식이 어떤 과(Department)을 위한 것인지 알 수 있어, "Past Medical History"가 특정 과에서 어떤 TDE들과 관계를 가지는지에 대한 정보를 얻을 수 있게 되었다.

5. 결론

본 연구에서는 임상 서식의 작성 행위에 내재된 지식을 태깅하는데 사용할 새로운 지식 모델을 제안하고, 이를 기반으로 CDT 온톨로지를 정의하였다. 또한, CDT 온톨로지를 이용한 태깅의 유용성을 확인하기 위해서 3명의 전문가가 참여하여 수작업으로 임상 서식을 태깅하고 그 결과를 지식베이스로 구축하였다.

구축된 지식 베이스를 살펴보면, 임상서식의 배치정보를 통해서 서식 항목 TDE가 사용된 목적과 환경에 따라 다른 TDE와 의미적으로 연결되어 있음을 볼 수 있다. 이러한 지식 베이스는 전문가가 아닌 사람에게도 임상서식에 내재된 CDT와 TDE에 대한 지식을 알기 쉽게 제공할 수 있다는 점에서 의미 있다.

Ⅵ. 임상 서식 지식베이스 기반의 서식 작성 지원 시스템

1. 시스템 소개

본 장에서는 5장에서 구축한 임상 서식 지식베이스를 이용하여 개발한 서식 작성 지원 시스템, STEP (Smart Clinical Document Template Editing and Production System)을 소개한다. STEP은 웹 기반의 지식베이스 관리 시스템으로 MVC (Model, View, Control) 패턴을 사용하여 개발되었으며, http://stem.bik.re.kr 에서 사용할 수 있다.

STEP 시스템은 5장에서 구축한 지식베이스를 통해 기존 EMR 시스템의 서식 작성과 서식 검색을 개선시키기 위해 개발되었다. 즉, 서식 작성자가 EMR 시스템에서 서식을 작성할 때 STEP이 보유한 지식을 제공하여, 작성자가 목적에 부합하는 서식을 더욱 빠르고 편리하게 작성할 수 있는 기능을 제공한다. 또한, TDE에서 사용되는 표준화되어 있지 않은 용어들을 SNOMED-CT와 같은 표준 용어와 연계할 수 있도록 하여 서식을 통한 정보 검색 및 관리가 일관성 있도록 지원한다. 이를 위해 STEP은 EMR시스템이나 용어관리 시스템과 같은 외부 시스템과의 유연한 연계를 고려하여 설계되었다.

STEP 시스템은 외부 시스템과의 연계 기능과는 별도로, 웹 인터페이스를 제공하여 사용자가 웹을 통해 지식베이스를 관리할 수 있도록 하였다. 사용자는 로그인을 한 후 시스템을 사용할 수 있으며, CDT와 TDE에 대한 정보를 시각화된 형식으로 제공받을 수 있다.

2. 시스템 구성

STEP 시스템은 Figure 28과 같이 크게 지식베이스의 관리를 위한 지식베이스 관리 모듈, 사용자와 외부 시스템 연계를 위한 웹 사용자 인터페이스, Web Services 인터페이스, 그리고 사용자의 요구를 처리하기 위한 핵심 모듈로 구성되었다.

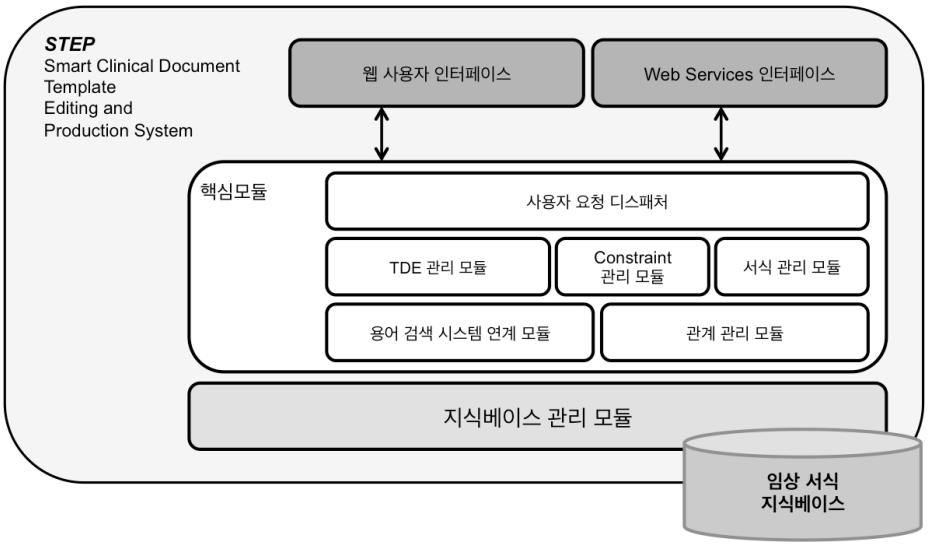

Figure 28 STEP 시스템의 구성

핵심 모듈과 웹 사용자 인터페이스 모듈간에는 데이터 교환 분야에서 주로 사용되는 JSON[13](JavaScript Object Notation)을 사용하여, 데이터를 보내고 받아서 처리하는 비용을 효율적이게 하였다. STEP에서 관리하는

[13] https://en.wikipedia.org/wiki/JSON

시스템은 서식 작성이 포함된 EMR 시스템과의 연계가 필수적인데, 이를 위해서 STEP 은 주요 기능을 SOAP(Simple Object Access Protocol)(W3C 2007. 27. April) 기반의 Web Services(w3c 11. Feb. 2004)로 제공한다. 다음은 3개의 주요 모듈과 각 모듈이 포함하는 세부 모듈에 대한 설명이다.

2.1. 지식 베이스 관리 모듈

지식 베이스는 임상 서식을 태깅하여 만들어진 온톨로지 인스턴스들로 구성되었다. 지식 베이스의 관리 모듈은 이들 인스턴스들에 대한 생성, 삭제, 검색, 그리고 변경을 위한 기능을 제공하며, 외부 모듈은 이 기능을 통해서 지식 베이스를 사용할 수 있도록 하였다. 지식 베이스 관리 모듈은 OWL 형식의 인스턴스들을 관리하기 위해 트리플 저장소, JENA TDB[14]를 사용하였으며, 이와 동일한 정보를 RDBMS에도 저장 관리하도록 하였다. 이렇게 동일한 정보를 두 가지 형식으로 저장관리 하는 이유는, 온톨로지를 이용한 추론 및 검색의 장점과 RDBMS의 안정된 관리 기능을 동시에 이용하기 위함이다.

2.2. 핵심 모듈

핵심 모듈은 6개의 세부 모듈로 구성되며, STEP 의 주요 기능을 제공한

[14] https://jena.apache.org/

다. 세부 모듈로는 사용자 요청 디스패처, TDE 관리 모듈, 서식 관리 모듈, Constraint 관리 모듈, 관계 관리 모듈, 그리고 용어 검색 모듈로 구성된다.

2.2.1. 사용자 요청 디스패처

사용자 요청 디스패처는 웹 사용자 인터페이스와 Web Services 인터페이스로 오는 요청을 전달될 모듈에 맞는 형식으로 변환하고 전송하는 역할을 한다. 즉, 사용자의 요청을 TDE 관리 모듈, 서식 관리 모듈, 관계 관리 모듈, 용어 검색 모듈에 맞게 변환하여 각각의 모듈이 제공하는 API (Application Programming Interface)에 전달하고, 그 결과를 JSON 형식으로 변환하여 반환한다. Web Services 인터페이스에서 호출되는 API는 외부 시스템과의 표준화된 연계를 위해 WSDL [15](Web Service Definition Language) 을 이용하여 정의한 후에 배포하였는데, 이렇게 한 이유는 연계 시스템에서는 WSDL를 이용하여 자동으로 연계 프로그램을 생성하고 사용할 수 있기 때문이다.

2.2.2. TDE 관리 모듈

TDE 관리 모듈은 서식 CDT에 포함된 서식 항목 TDE를 생성, 삭제, 검색, 그리고 수정할 수 있는 기능을 제공한다. TDE에 대한 정보는 이름과 TDE의 Constraint에 해당하는 ValueSet, SNOMED-CT와 같은 외부 자원과의 연계를 위한 reference, 그리고 서식 CDT나 서식 항목 TDE와의 관

[15] http://www.w3.org/TR/wsdl

계 정보를 포함한다. TDE 관리 모듈은, 이러한 기능들을 구현하기 위해 지식베이스 관리 모듈과 관계 관리 모듈에서 제공하는 기능을 이용한다.

2.2.3. 서식 관리 모듈

서식 관리 모듈은 서식을 생성하고 관리하는 기능과 TDE와의 관계를 관리하는 기능을 제공한다. TDE 와의 관계는 관계 관리 모듈에서 관리되며, 이를 통해 TDE가 어떤 서식에서 사용되었는지에 대한 정보를 얻을 수 있다.

2.2.4. Constraint 관리 모듈

STEP에서 TDE가 가질 수 있는 값의 범위는 Constraint에 의해서 결정되는데, Constraint 관리 모듈은 이러한 값의 범위를 표현하는 ValueSet을 관리한다. Constraint 모듈은 값의 범위를 생성, 삭제, 수정, 그리고 검색할 수 있는 기능을 제공할 뿐 아니라, SNOMED-CT나 ICD-10과 같은 외부 용어 체계와 연동하여 값의 범위를 정의할 수 있다. 이러한 기능은 용어 검색 시스템 연계 모듈과 관계 관리 모듈을 사용한다.

2.2.5. 관계 관리 모듈

새로운 TDE가 하나 만들어지는 것은 여러 정보의 추가적인 생성을 의미한다. TDE가 생성되면, 먼저 TDE가 사용된 서식과 관계가 만들어져야 하며, TDE의 Constraint에 해당하는 ValueSet과 연결하는 것, 그리고 필요하다면 TDE이름과 외부 참조 용어가 연결되어야 것, 마지막으로 다른 TDE

와의 관계가 설정되어야 한다. 관계 관리 모듈에서는 이 과정에서 생성되는 관계 정보를 관리한다.

2.2.6. 용어 검색 시스템 연계 모듈

이 모듈은 외부 용어 관리 시스템에서 제공하는 검색 기능을 사용하기 위한 것이다. STEP 시스템은 서울대학교 치과대학 의생명지식공학 연구실에서 개발한 용어 관리 시스템인 LexCare Suite(Lee, Song et al. 2010)과 연계되어 있으며, LexCare Suite에서 제공하는 SNOMED-CT와 ICD-10 검색 기능을 사용한다. LexCare Suite은 LexGrid Model(Pathak, Solbrig et al. 2009)을 사용하여 표준 용어 체계가 가지는 다양한 정보 구조에 영향을 받지 않고 용어의 검색과 용어 체계간 맵핑을 정의하고 관리할 수 있는 기능을 제공한다. LexCare Suite은 표준 용어 체계뿐 아니라, 병원에서 사용하는 표준화 되지 않은 용어를 표준 용어 체계와 동일하게 검색하고 관리할 수 있는 기능을 제공한다.

용어 검색 시스템 연계 모듈은 LexCare Suite에서 제공하는 Web Services를 사용하여, STEP의 모듈들이 LexCare Suite가 제공하는 용어 검색 기능을 마치 한 시스템에서 제공하는 것처럼 사용할 수 있도록 한다. STEP은 TDE의 이름과 ValueSet을 정의하는 과정에서 표준 용어 체계를 검색할 수 있는 기능을 제공한다. 이 때 TDE의 이름을 표준 용어 체계와 연결하는 기능은 매우 중요한데, 서식에서는 동일한 의미의 단어에 대해서도 여러 가지 표현을 사용하기 때문이다. 예를 들어, TDE "Past Illness"는 "Past Medical History"나 "과거력"등과 같은 의미지만, 기존의 EMR에서

는 문자열의 차이로 다르게 처리된다. 그런데 각각의 TDE가 SNOMED-CT의 " Past medical history"으로 연결되어 있다면, 위의 세 TDE는 동일한 것으로 처리되어 용어간의 의미적 상호운용성을 확보할 수 있을 것이다.

용어 검색 시스템 연계 모듈과 LexCare Suite가 Web Services 기술을 통해 연결된 것은 여러 가지 장점을 제공한다. 먼저, LexCare Suite의 변경이나 개선에도 STEP 시스템의 변경 없이 개선된 기능을 사용할 수 있으며, LexCare Suite외에 다른 시스템과도 연계될 수 있는 유연성을 가질 수 있다. STEP 과 외부 시스템과의 연계는 Figure 29 과 같다. 그림에서는 용어서버 이외에도 STEP에서 제공하는 Web Services를 사용하는 서식 작성기와의 연계도 포함하였다.

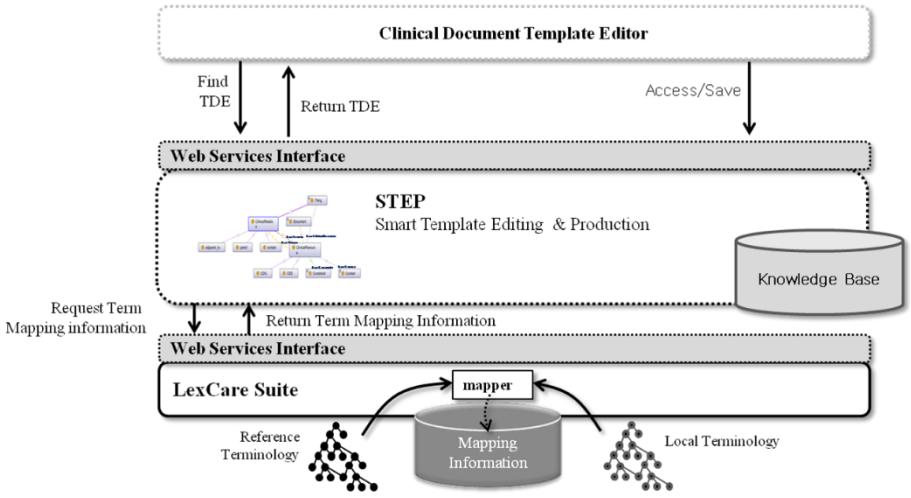

Figure 29 STEP 시스템과 외부 시스템과의 연계

2.3. 웹 사용자 인터페이스

웹 기반의 사용자 인터페이스는 HTML5와 JQuery를 이용하여 개발되었다. 사용자 인터페이스는 사용자 클릭과 같은 이벤트를 처리하기 위해 핵심 모듈에 요청을 보내며, 그 결과로 온 JSON 형식의 데이터를 사용하여 화면을 생성한다. 이러한 과정을 통해 사용자는 CDT와 TDE를 조회할 수 있다.

Figure 30은 사용자가 STEP을 사용할 때, 가장 먼저 볼 수 있는 화면으로 사용자 로그인 화면이다.

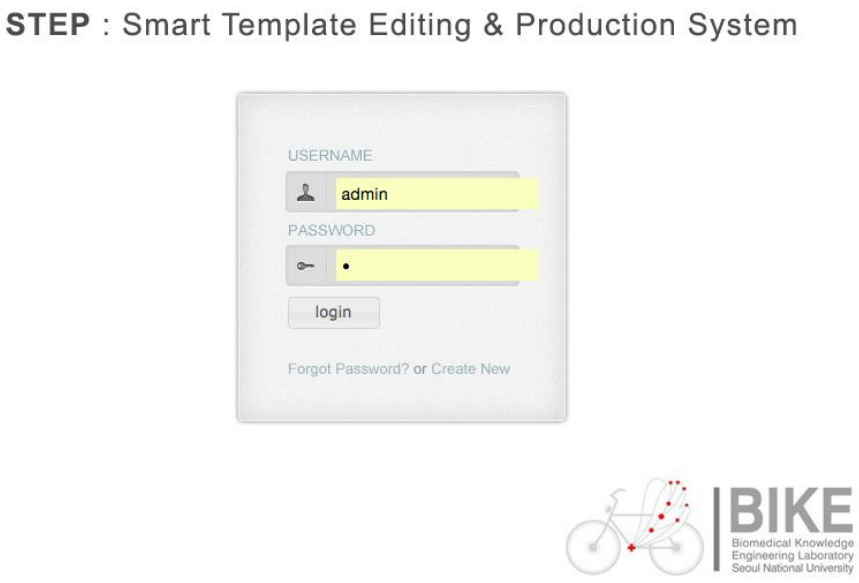

Figure 30 STEP 로그인 화면

Figure 31은 STEP 시스템에서 관리하는 서식과 서식의 정보를 볼 수 있는 화면이다. 화면의 왼쪽을 보면 관리하고 있는 서식 리스트를 볼 수 있으

며, 사용자가 선택한 서식의 정보는 오른쪽에서 볼 수 있다. 서식의 정보는 이름과 ID로 구성된 기본정보와 서식에 포함된 TDE 리스트로 구성된다. 사용자는 서식에서 사용된 TDE 이름을 클릭하여 관련된 정보를 확인할 수 있다.

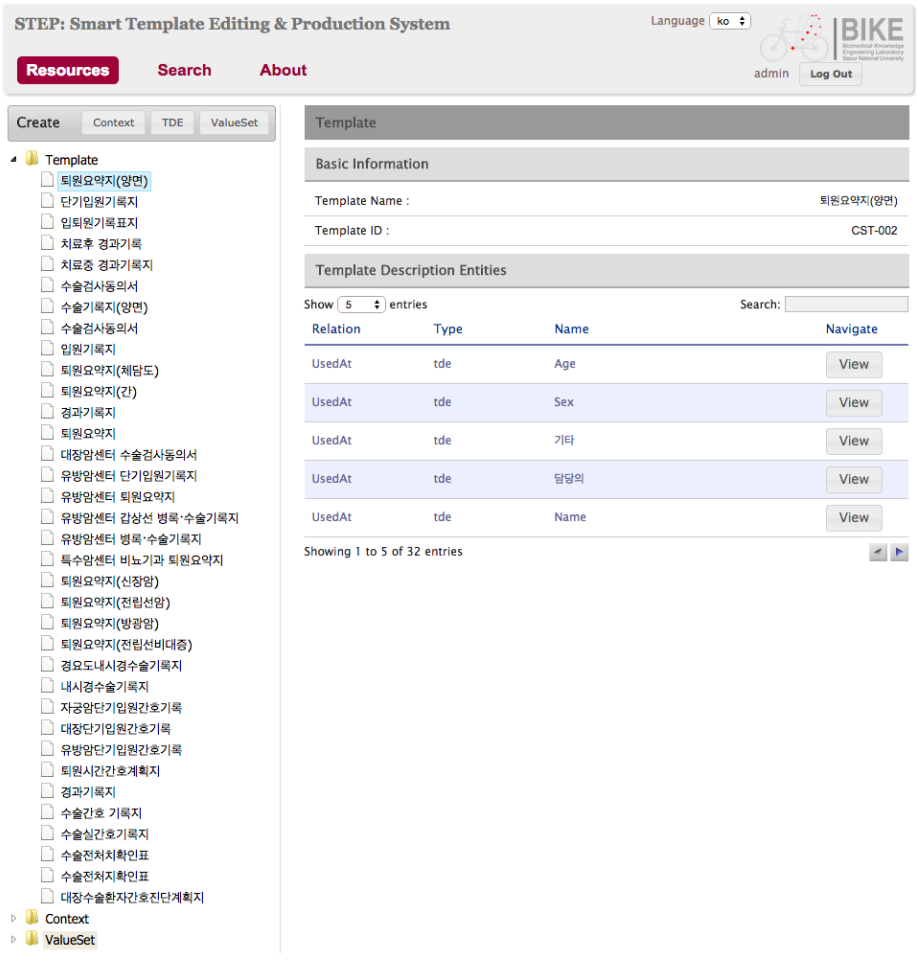

Figure 31 STEP에서 임상 서식 관리

Figure 32는 웹 인터페이스에서 TDE를 관리하기 위한 화면이다. 왼쪽

화면은 TDE들을 TDE가 사용된 Context에 맞게 구분한 것이며, 오른 쪽 화면은 TDE를 선택했을 때 보여지는 화면이다. Figure 32은 그 중에서도 TDE "Past Illness"를 선택한 화면이다.

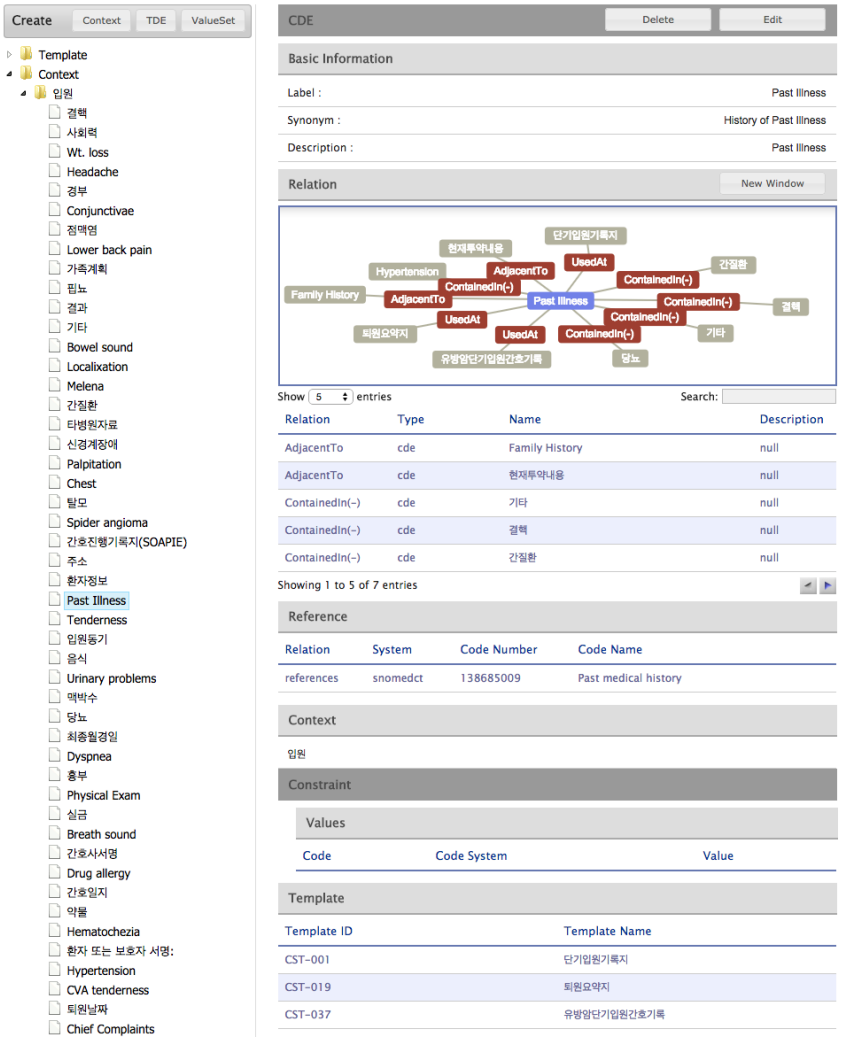

Figure 32 STEP에서 TDE 관리

TDE에 대한 정보는 이름과 설명, 그리고 유사어를 포함하는 기본 정보

와 관계정보, 외부 용어와의 참조, Context, Constraint, 그리고 TDE가 사용된 서식 리스트를 보여준다. 먼저 관계 정보에서는 TDE "Past Illness"와 관계를 맺고 있는 다른 TDE들을 그래프와 테이블로 보여주고 있다. 또한 외부 용어와의 참조를 위한 Reference에서는 이 "Past Illness"가 SNOMED-CT의 "Past medical history"와 맵핑되었음을 보여주고 있다. 마지막으로 Template 항목을 통해, 이 TDE가 "단기입원기록지", "퇴원요약지", 그리고 "유방암단기입원간호기록"에서 사용되었음을 알 수 있다.

Figure 33 은 TDE의 Constraint로 연결될 수 있는 ValueSet을 관리하기 위한 화면이다. 그림에서 "횟수_정수" ValueSet은 SNOMED-CT에서 정의한 "Positive Integer"값을 가지도록 정의되었다. 따라서 이 ValueSet을 Constraint로 사용하는 TDE는 양의 정수 값을 가져야 하며, 그렇지 않은 경우 유효하지 않은 상태가 된다.

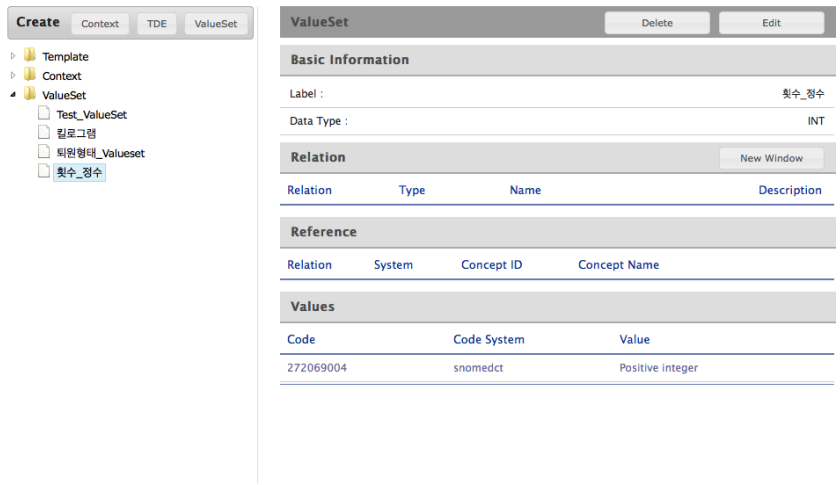

Figure 33 STEP에서 ValueSet의 관리

마지막으로, Figure 34은 LexCare Suite에서 제공하는 용어검색 기능을 이용하는 것으로, TDE 를 편집할 때 볼 수 있는 화면이다. 현재 LexCare Suite는 SNOMED-CT와 ICD-10에 대한 검색을 제공하고 있으며, 화면은 SNOMED-CT를 대상으로 "secondary hypertension"을 검색한 결과이다. 만약 사용자가 결과 중에서 6번째 결과인 "Secondary hypertension"을 선택하였다면, STEP은 해당 용어의 ID와 기본적인 정보, 그리고 화면의 아래쪽에서 선택한 관계 정보를 TDE와 연관 지어 저장한다. 사용자는 다양한 관계를 정의할 수 있으나, 현재는 "references"만 정의하였다.

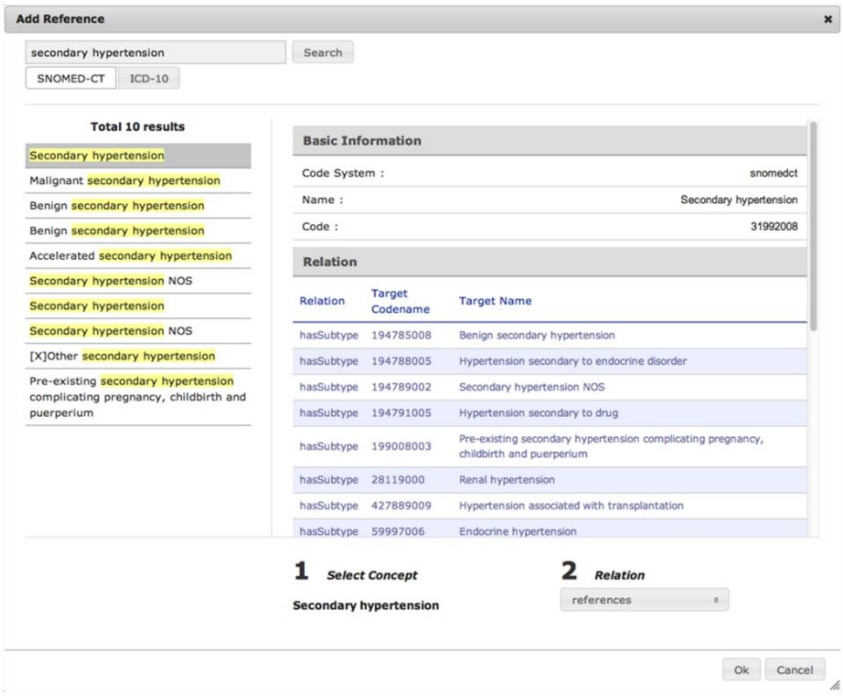

Figure 34 용어검색 시스템에서 "Secondary hypertension" 검색

2.4. Web Services 인터페이스

Web Services 인터페이스는 STEP이 보유한 지식 베이스를 서식 편집기와 같은 외부 시스템에서도 사용할 수 있도록 유연한 연계 기능을 제공한다. 이를 위해 본 모듈에서는 총 13개의 SOAP기반의 웹 서비스 오퍼레이션을 제공하며, 각각의 오퍼레이션은 제공하는 기능에 따라 Get, Find, 그리고 Save 접미사가 부착되었다. Table 22은 각각의 오퍼레이션에 대한 설명이다.

Table 22 STEP의 웹 서비스

오퍼레이션 이름	설명
getTDEListByContextID	사용자가 지정한 Context ID에 해당하는 서식 항목 TDE 리스트를 반환한다.
getContextList	STEP에 저장된 모든 Context 리스트를 반환한다.
getTemplateList	STEP에 저장된 모든 임상서식 CDT 리스트를 반환한다.
getTDEListByTemplateID	사용자가 지정한 임상 서식에 포함된 모든 서식 항목 TDE 리스트를 반환한다.
getTDEInfo	사용자가 지정한 서식 항목에 해당하는 서식 항목 TDE의 상세 정보를 제공한다.
findTDE	사용자의 검색 키워드와 유사한 서식항목 TDE 리스트를 반환한다. 사용자는 TDE가 사용된 Context나 임상서식을 지정하여 검색을 한정시킬 수 있으며, 파라미터 limit과 offset을 이용하여 검색 결과 중 일부분만을 가져올 수 있다.
findRelatedTDE	사용자가 지정한 서식항목 TDE와 유사한 TDE 리스트를 반환한다. 사용자는 TDE가 사용된 Context나 임상서식을 지정하여 검색을 한정시킬 수 있으며, 파라미터 limit과 offset을 이용하여 검색 결과 중 일부분만을 가져올 수 있다.

findTemplate	사용자가 지정한 서식 항목 TDE를 포함하는 모든 임상 서식 CDT리스트를 반환한다. 사용자는 파라미터 limit과 offset을 이용하여 검색 결과 중 일부분만을 가져올 수 있다.
findRelatedTemplate	사용자가 지정한 임상서식 CDT와 유사한 CDT 리스트를 반환한다. 사용자는 파라미터 limit과 offset을 이용하여 검색 결과 중 일부분만을 가져올 수 있다.
saveTDE	서식 항목 TDE를 저장한다.
saveRelation	관계정보를 저장한다.
saveTemplate	임상서식 CDT를 저장한다.
saveValueSet	ValueSet를 저장한다.

Figure 35은 STEP의 Web Services 중 findTDE가 호출되었을 때 요청/응답 메시지를 보여주고 있다. 예에서, 사용자는 "past"란 키워드를 사용하여, STEP이 보유한 TDE 중에서 이름에 "past" 문자열을 포함하고 있는 것들을 찾고자 하였으며, 검색 결과는 10개로 한정 시켰다. STEP의 응답 메시지를 보면, 사용자 검색 요청을 만족하는 TDE "Past Medical History"와 "Past Illness"가 반환되는 것을 볼 수 있다

Figure 35 findTDE 웹 서비스의 Request/Response 메시지

3. Use Case

본 연구에서는 임상 지식 베이스의 활용성을 검증하기 위해 STEP에서 제공하는 Web Services을 이용하여 GUI기반의 임상 서식 편집기를 개발하였다. 서식 편집기의 모습은 Figure 36과 같다. 서식 편집기는 5개의 패널로 구성되는데, 첫 번째는 STEP에서 제공하는 모든 TDE를 Context를 기준으로 보여주기 위한 TDE 리스트 패널이다. 이 패널에서 사용자는 Context를 기준으로 TDE를 탐색하거나 검색 기능을 이용하여 TDE를 검색할 수 있다. 두 번째 패널은 편집되는 서식을 볼 수 있는 미리보기 패널로 사용자가 TDE를 선택하는 경우 이를 반영하여 미리 서식의 모습을 볼 수 있는 기능을 제공한다. 세 번째는 사용자가 선택한 TDE 리

스트를 볼 수 있는 사용자 패널이다. TDE 리스트 패널에서 선택한 TDE를 드래그/드롭하는 경우, TDE는 사용자 패널에 추가된다. 이 패널에서 사용자는 선택된 TDE를 삭제 시키거나 현재 리스트를 이용하여 미리 보기 기능을 실행시킬 수 있다. 네 번째 패널은 오른쪽 위에 있는 임상 서식 패널로 사용자가 선택한 TDE가 사용된 서식 리스트를 보여준다. 사용자는 서식 리스트 중에 있는 서식을 선택함으로 기존의 서식이 어떤 모습인지 확인할 수 있으며, 복사를 통해서 새로운 서식을 만들 수도 있다. 마지막, 다섯 번째 TDE 추천 패널은 사용자가 선택한 서식 항목과 관계 있는 TDE 리스트를 제공한다. 이러한 정보는 서식을 설계 하는데 있어 기존 서식이 보유한 전문 지식을 재사용할 수 있게 하는 것으로, 새로운 서식을 작성하는데 있어 가치 있는 정보라 할 수 있다. 또한 사용자는 Context와 사용된 과(Department)를 한정시키며 다양한 관점의 정보를 필터링하여 제공받을 수 있다.

Figure 36 에서는 임상 서식기에서 제공하는 기능을 이용하여, 사용자가 서식을 편집하는 과정을 보여주고 있다. 그림에서 사용자는 "admission note"란 이름의 서식을 만들고 있으며, TDE "patient information"과 " social history"을 왼쪽 패널의 TDE 리스트에서 드래그 드롭하여 사용자 패널에 추가하였다. 만약 사용자 패널에서 특정 TDE가 선택되면, TDE 추천 패널에서는 이와 관련된 TDE를 보여 주는데, 화면에서는 사용자가 "Social History"를 선택하였다. 이에 따라 추천 패널에서는 "Social History"와 "ContainedIn" 관계를 가진 "blood type"과 "weight", 그리고 "교육 수준"와 같은 TDE를 추천하였으며, 이외에도

"AdjacentTo"관계로 연결된 "Past Medical History"를 추천하였다. 사용자는 추천된 TDE를 드래그/드롭하여 사용자 패널에 추가 시킬 수 있는데, 이런 과정은 기존의 서식 작성 지식을 활용하는 예가 될 수 있다. 이런 과정을 통해 작성된 서식은 문서 형식으로 저장될 수 있으며, 다시 STEP에 새로운 임상 서식으로 저장될 수 있다.

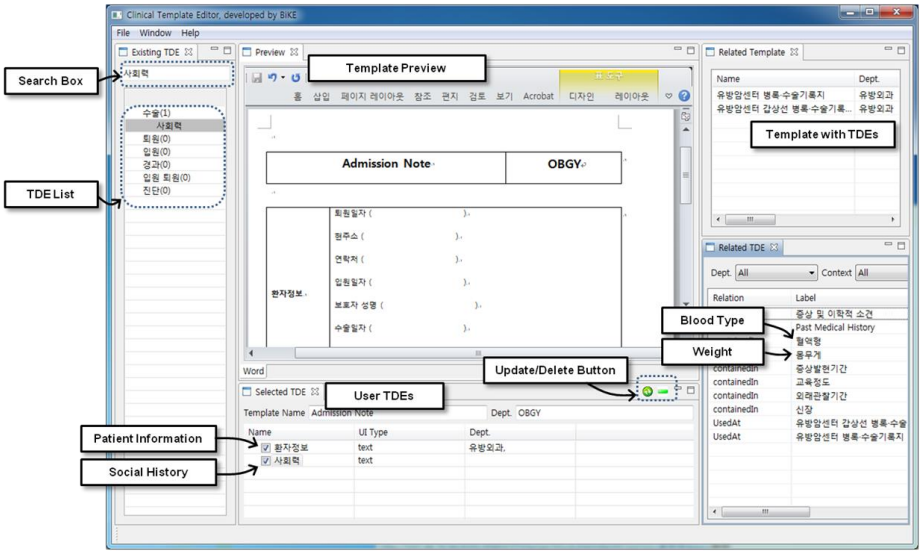

Figure 36 STEP을 이용한 임상 서식 편집기

4. 결론

본 연구에서 제안한 STEP 시스템이 기여하는 바는 두 가지이다. 첫 번째 STEP은 임상 서식에 내재된 지식을 활용하여 새로운 서식 과정이 보다 지능적인 방법으로 개선될 수 있음을 보여준다. STEP이 제안한 지식 모델은 임상 서식의 표현 방식에 독립적인 모델로, 기존의 임상 개념 모델 (Beale 2003, Goossen, Goossen-Baremans et al. 2010)과는

차이가 있다. 기존의 연구가 임상 개념의 표현을 위한 구조와 의미를 정의하고자 하였다면, STEP은 임상 서식이 어떻게 만들어 지고, 그 안에 어떤 서식 항목들이 사용되었는지에 초점을 맞추어, 임상 개념을 표현하는 모델이 어떤 것이 사용되었는지에 대해서는 영향을 받지 않는다. 단, STEP의 지식 모델은 기존 임상 개념 모델과 같이 EMR 시스템에 적용되기 위해서는 더욱 구체적인 연계 구조가 필요하며, 이를 위해서는 외부 임상 모델 연구와의 협업이 필수적이다.

참조 용어는 일반적으로 질병이나 치료와 같은 특정 도메인을 대상으로 작성된다. 특정 도메인에 한정되지 않은 임상 개념 모델은 이러한 참조 용어를 사용하여 개념의 구조를 조직화한다. 그러나, STEP의 지식 모델은 기존 모델들과는 달리 실용적인 접근을 취한다. 즉, 임상 서식에서 사용될 수 있는 것이라면, 그 것이 단순 개념이던지, 아니면 여러 개념으로 구성된 복잡한 개념이던지 상관없이 사용할 수 있다. 이러한 실용적인 접근이 가능했던 것은, 온톨로지를 통해 기존 임상 개념 모델과 참조 용어를 이용하여 새로운 개념을 표현하는 방식을 선택했기 때문이다.

두 번째, STEP은 임상 서식과 표준 용어를 어떻게 연결시킬 것인가에 대한 실용적인 대안을 제시하였다. 대학 병원에서 사용하는 용어는 진료과목과 업무의 특성에 따라 다양한 변이를 가지며 사용될 수 밖에 없는데, 이러한 현상은 같은 개념이면서도 다르게 표현되는 수 많은 변이의 원인이 된다. STEP은 이러한 문제점을 임상 서식 수준에서 해결할 수 있는 실제적인 방안을 제시하였다.

STEP이 비록 웹 버전과 임상 서식 에디터를 통해 POC(Proof of

Concept)을 하였으나, 여전히 개발의 여지가 있다. STEP은 다음 3가지 방향에서 지속적인 연구와 개발이 필요하다. 첫 번째, STEP에서 보유한 임상 서식의 증가에 따라, 임상 서식을 그룹핑 하거나 검색할 때에 필요한 새로운 서식간 유사도 계산 방법이 필요하다. STEP의 모든 TDE와 CDT는 온톨로지로 표현되어 있기 때문에 이들간의 유사도 계산은 새로운 방법이 필요하다. 향후 의생명 분야 온톨로지간의 유사도 계산과 관련된 연구(Pesquita, Faria et al. 2009)를 기반으로 STEP 또한 새로운 계산 방식이 적용되어야 한다. 두 번째, STEP과 LexCare Suite에서 제공하는 Web Services를 용어 서비스와 관련된 국제 표준인 CTS[16](Common Terminology Services)에 호환되도록 개선하여, STEP의 CDT와 TDE를 표준화된 방법으로 접근할 수 있도록 해야 한다. 세 번째, STEP의 지식모델에 포함된 임상 서식의 출처 정보를 이용하는 연구가 필요하다. 의생명 분야와 시맨틱 웹 분야에서는 최근, 시간에 따른 데이터의 변화와 이에 따른 새로운 검색 방법이 연구되고 있는데(Perry, Jain et al. 2011, Hoffart, Suchanek et al. 2013, Sun, Rumshisky et al. 2013), STEP에서도 임상서식과 서식항목을 포함된 출처 정보를 사용하도록 하여 시간에 따른 검색이 가능하도록 할 뿐 아니라 임상 현장에서 Copy&Paste로 발생하는 잘못된 정보의 확산(Hammond, Helbig et al. 2003)을 방지할 수 있어야 한다.

---

[16] http://www.omg.org/spec/CTS2/1.1/

Ⅷ. 결론

본 연구는 의생명 분야의 대표적인 문서인 연구 논문과 임상 문서에 대해 사용자의 다양한 정보 접근의 요구를 만족시키기 위한 것이다. 이러한 목적을 위해서, 본 연구에서는 각각의 문서에 적합한 어노테이션 방법을 제안하고 이를 자동화하고 지식화하는 연구를 하였다.

먼저, 텍스트 형식의 연구 논문에 대해서는 연구 활동의 방향 설정에 중요한 역할을 하는 초록을 대상으로, 의생명 분야에서 주로 사용하는 IMRAD로의 자동 태깅을 시도하였다. 이 연구에서는, 기존 언어학 분야에서 의생명 분야의 논문을 대상으로 이룬 결과와 컴퓨터 과학 분야에서 진행되어 온 결과를 살펴보고, 두 분야의 연구 성과를 이용하여 계산 비용이 적으면서도 높은 성능을 내는 자동 태깅 시스템을 개발하였다. 본 연구에서 제안한 방법을 사용하는 경우, 문장에서 뽑아낸 17개의 특징만으로도 비구조화된 초록을 Accuracy 77.0 ~ 90.3%의 성능으로 분류할 수 있음을 보여주었고, 기존 특징과 조합했을 때는 최고 Accuracy 91.7%의 성능을 보여주었다. 이러한 실험 결과는 본 논문에서 사용한 의생명 분야의 언어적 특징이 문장 분류에서 좋은 특징으로 사용될 수 있음을 보여준다. 또한, 언어적 특징은 의생명 분야에서 높은 정보 가치를 가지고 있다는 RCT 초록을 대상으로도 좋은 성능을 보여주었다. RCT 초록을 대상으로 언어적 특징을 사용했을 때, 기존 BOW를 사용했을 때 보다 최저 3.7%에서 최고 35.8%의 성능 향상을 보였다.

임상 문서의 경우, EMR을 시스템을 사용하는 환경에서는 임상 서식을

통해 생성되는 경우가 대부분이므로, 임상 서식을 대상으로 자동 태깅을 시도하였다. 임상 서식은 연구 초록과는 달리 이미 구조화된 형식을 가지고 있으므로, 본 연구에서는 이 구조 안에 있는 내재된 전문가의 지식을 태깅하고자 하였다. STEP은 이러한 연구의 결과물로, 임상 서식에 내재된 전문가의 지식을 위한 지식모델을 정의하고 이를 통해 지식베이스를 포함하며, 이러한 지식베이스가 EMR시스템에서 사용될 수 있도록 표준적인 연계방법을 제공하였다. STEP에서 사용하는 지식 모델은 임상 서식에서 태깅된 정보를 지식화하기 위한 수단으로 두 가지 측면에서 기여하는 바가 크다. 첫 번째, 병원에서 일상적으로 사용하는 임상 서식을 병원의 지식으로 축적시킬 수 있고, 이를 통해 임상 서식 작성 과정을 개선시킬 수 있음을 보여주었다. 두 번째, STEP은 임상 서식과 표준 용어를 어떻게 연결시킬 것인가에 대한 실용적인 대안을 제시하였다. 병원에서 사용하는 임상 서식에서부터 용어 표준을 시작하는 방법은 기존 임상 개념 수준의 방법보다 실제적인 접근일 수 있다.

의생명 분야의 대표적인 문서들에 대한 이와 같은 연구는, 두 가지 점에서 중요하다. 첫 번째, 본 연구는 의생명 분야의 대표적인 두 문서의 검색과 관리가 효율적으로 개선될 수 있음을 보여준다. 급속도로 증가하는 두 문서의 특성상 문서의 정확한 검색과 효율적인 관리는 점점더 중요한 이슈로 인식되고 있는데, 본 연구의 결과는 이러한 이슈를 해결할 수 있는 기반 기술을 제공한다는 점에서 중요하다. 두 번째, 본 연구 결과는 의생명 분야의 텍스트 마이닝 연구나 응용 프로그램 개발에 도움이 되는 기반 기술을 제공한다. 의생명 분야의 중요성이 인식되면서 최근 의생명 문서

를 분석하는 연구나 정보 시스템이 활발히 이루어지고 있는데, 본 연구의 결과는 이들 연구에서 활용할 수 있는 실제적인 결과물을 제시한다는 점에서 의미있다. 이러한 맥락에서 본 연구에서는 연구 과정에서 추출한 자원과 구축한 최종 시스템을 웹에 공개하여 다른 연구자들에게 활용될 수 있도록 하였다. 저자는 이렇게 공개된 자원과 시스템이 국내외 의생명 분야의 발전에 기여하길 바란다.

Ⅷ. 연구의 제한점 및 제언

본 연구의 과정과 결론을 통해 다음과 같이 몇 가지 한계점과 새로운 연구 방향을 제안하고자 한다.

첫째, 비구조화된 초록의 문장을 태깅하기 위해서는 본 연구에서는 SVM 알고리즘을 사용하였다. SVM은 문장 분류에 많이 사용되어 온 알고리즘으로 성능상으로도 좋은 점이 있으나, 최근에는 CRF (Conditional Random Field)나 SVM struct 와 같은 알고리즘이 좋은 성능을 보여주고 있다 (Keerthi and Sundararajan 2007). 향후 문장 태깅의 성능을 개선하기 위해서는 이와 같은 새로운 알고리즘의 채택이 필요하다.

둘째, 비구조화된 초록의 자동 태깅이 실제 환경에서 의미있게 활용되기 위해서는 대규모 문서 처리에 대한 고려가 필요하다. 본 연구에서는 단일 노드에서 사용할 수 있는 Weka를 사용하였는데, Apache Mahout와 같이 대규모 문서를 대상으로 확장 가능한 성능을 보여주는 공개 소프트웨어를 도입하는 것도 좋은 대안일 수 있다.

셋째, 비구조화된 초록의 자동 태깅이 더욱 의미있게 활용되기 위해서는, IMRAD로 구별된 문장들에서 정보를 추출하는 연구가 연속되어 이루어져야 한다. 예를 들어, Introduction 섹션의 문장들에서 연구 목적과 관련된 문장만을 필터링하여 연구 목적과 관련된 정보만을 추출하여 구조화한다면 논문 검색과 분석은 더욱 개선될 수 있을 것이다.

넷째, STEP시스템에서 제안한 임상 서식을 위한 지식 모델은 비록 POC(Proof Of Concept)를 통해서 유용성을 검증하였지만, 병원 시스템과의 연동을 통한 실제적인 평가는 이루어지지 않았다. STEP 시스템이

제안한 지식 모델이 병원 문서 관리에 도움을 주기 위해서는, 실제 EMR 시스템과 연계되어 사용자의 다양한 검색 요구가 지식 모델에 반영되어야 한다.

다섯째, 저자는 본 연구의 결과물들이 다양한 연구 분야에서 활용되기를 기대한다. 비구조화된 초록을 IMRAD로 태깅하는 연구 결과는, 논문의 초록을 분석하여 논문의 트렌드를 분석하거나 논문 검색의 정확률을 높이려는 연구에서 중요하게 활용될 수 있을 것이다. 논문 탐색과 동향 파악은 연구자들이 연구 과정에서 일상적으로 하는 것으로, 이러한 활동은 구조화된 초록으로 태깅된 논문들로 인해 더욱 효율적으로 개선될 수 있을 것이다. STEP 시스템에서 제안한 지식 모델인 CDT 온톨로지 역시, 웹으로 공개하여 임상 서식의 지식화가 필요한 기관에서 사용할 수 있도록 하였다. 본 연구에서 제안하고 개발한 시스템은 위에서 언급한 대로 한계점이 있으나, 지식 모델의 도입만으로도 병원 서식의 관리나 표준화 도입에 기여할 수 있을 것이다.

여섯째, 모든 연구 결과물들은 웹에 공개하여 관련 연구에 기여하고자 하였다. 비구조화된 초록의 태깅에 대한 결과는 http://abstract.bike.re.kr 에, STEP 시스템은 http://step.bike.re.kr 에서 확인할 수 있다. 이러한 정보 공개가 의생명 분야의 발전에 기여하기를 바란다.

참고문헌

Ananiadou, S. and J. McNaught (2006). Text mining for biology and biomedicine, Citeseer.

Batagelj, V. and A. Mrvar (2002). Pajek—analysis and visualization of large networks. Graph Drawing, Springer.

Bayley, L. and J. D. Eldredge (2003). " The structured abstract: an essential tool for researchers." Hypothesis **17**(1): 11-13.

Beale, T. (2003). "Archetypes and the EHR." Studies in health technology and informatics: 238-246.

Biber, D., S. Conrad and V. Cortes (2004). "If you look at …: Lexical Bundles in University Teaching and Textbooks." Applied Linguistics **25**(3): 371-405.

Blumenthal, D. and M. Tavenner (2010). "The “meaningful use” regulation for electronic health records." New England Journal of Medicine **363**(6): 501-504.

Bruza, P. and M. Weeber (2008). Literature-based discovery, Springer Science & Business Media.

Budgen, D., A. J. Burn and B. Kitchenham (2011). "Reporting computing projects through structured abstracts: a quasi-experiment."

Empirical Software Engineering: 1-34.

Chen, R., G. O. Klein, E. Sundvall, D. Karlsson and H. Åhlfeldt (2009). "Archetype-based conversion of EHR content models: pilot experience with a regional EHR system." BMC medical informatics and decision making **9**(1): 33.

Chung, G. (2009). "Sentence retrieval for abstracts of randomized controlled trials." BMC Medical Informatics and Decision Making **9**(1): 10.

Cohen, A. M. and W. R. Hersh (2005). "A survey of current work in biomedical text mining." Briefings in bioinformatics **6**(1): 57-71.

Cortes, V. (2013). "The purpose of this study is to: Connecting lexical bundles and moves in research article introductions." Journal of English for Academic Purposes **12**(1): 33-43.

Coyle, J., Y. Heras, T. Oniki and S. Huff (2008). "Clinical element model." University of Utah.

Coyle, J. F., A. R. Mori and S. M. Huff (2003). "Standards for detailed clinical models as the basis for medical data exchange and decision support." International journal of medical informatics **69**(2): 157-174.

Csomay, E. (2012). "Lexical Bundles in Discourse Structure: A Corpus-Based Study of Classroom Discourse." Applied Linguistics.

de Waard, A. and H. Pander Maat (2012). "Verb form indicates discourse segment type in biological research papers: Experimental evidence." Journal of English for Academic Purposes **11**(4): 357-366.

Druss, B. G. and S. C. Marcus (2005). "Growth and decentralization of the medical literature: implications for evidence-based medicine." Journal of the Medical Library Association **93**(4): 499.

Fleuren, W. W. and W. Alkema (2015). "Application of text mining in the biomedical domain." Methods **74**: 97-106.

Ganiz, M. C., W. M. Pottenger and C. D. Janneck (2005). "Recent advances in literature based discovery." Journal of the American Society for Information Science and Technology, JASIST (Submitted).

Garde, S., R. Chen, H. Leslie, T. Beale, I. McNicoll and S. Heard (2009). Archetype-based knowledge management for semantic interoperability of electronic health records. Mie.

Gerstein, M., M. Seringhaus and S. Fields (2007). "Structured digital abstract makes text mining easy." Nature **447**(7141): 142.

Goossen, W., A. Goossen-Baremans and M. van der Zel (2010). "Detailed clinical models: a review." Healthcare informatics research **16**(4): 201-214.

Group, T. S. o. R. T. (1994). "A proposal for structured reporting of

randomized controlled trials." The Journal of the American Medical Association **272**(24): 1926-1931.

Groza, T., H. Hassanzadeh and J. Hunter (2013). "Recognizing Scientific Artifacts in Biomedical Literature." Biomedical informatics insights **6**: 15.

Guo, Y., A. Korhonen, M. Liakata, I. S. Karolinska, L. Sun and U. Stenius (2010). Identifying the information structure of scientific abstracts: an investigation of three different schemes. Proceedings of the 2010 Workshop on Biomedical Natural Language Processing. Uppsala, Sweden, Association for Computational Linguistics**:** 99-107.

Hall, M., E. Frank, G. Holmes, B. Pfahringer, P. Reutemann and I. H. Witten (2009). "The WEKA Data Mining Software: An Update." SIGKDD Explorations. **11**(1).

Hammond, K. W., S. T. Helbig, C. C. Benson and B. M. Brathwaite-Sketoe (2003). Are electronic medical records trustworthy? Observations on copying, pasting and duplication. AMIA Annual Symposium Proceedings, American Medical Informatics Association.

Hanania, E. A. S. and K. Akhtar (1985). "Verb form and rhetorical function in science writing: A study of MS theses in biology, chemistry, and physics." The ESP Journal **4**(1): 49-58.

Harbourt, A. M., L. S. Knecht and B. L. Humphreys (1995).

"Structured abstracts in MEDLINE, 1989-1991." Bulletin of the Medical Library Association **83**(2): 190-195.

Hearst, M. A. (1999). Untangling text data mining. Proceedings of the 37th annual meeting of the Association for Computational Linguistics on Computational Linguistics, Association for Computational Linguistics.

Hirohata, K., N. Okazaki, S. Ananiadou and M. Ishizuka (2008). Identifying sections in scientific abstracts using conditional random fields. the Third International Joint Conference on Natural Language Processing (IJCNLP 2008), Hyderabad, India.

Hoffart, J., F. M. Suchanek, K. Berberich and G. Weikum (2013). YAGO2: A spatially and temporally enhanced knowledge base from Wikipedia. Proceedings of the Twenty-Third international joint conference on Artificial Intelligence, AAAI Press.

Huth, E. J. (1987). "Structured abstracts for papers reporting clinical trials." Ann Intern Med **106**(4): 626-627.

Hyland, K. (2008). "Academic clusters: text patterning in published and postgraduate writing." International Journal of Applied Linguistics **18**(1): 41-62.

Jeong, S., S. Nam and H.-Y. Park (2014). "An ontology-based biomedical research paper authoring support tool." Science Editing

1(1): 37-42.

Kang, N., E. M. van Mulligen and J. A. Kors (2011). "Comparing and combining chunkers of biomedical text." Journal of biomedical informatics **44**(2): 354-360.

Keerthi, S. S. and S. Sundararajan (2007). CRF versus SVM-struct for sequence labeling, Technical report, Yahoo Research.

Kim, H.-G., B.-H. Ha, J.-I. Lee and M.-K. Kim (2005). "A multi-layered application for the gross description using semantic web technology." International journal of medical informatics **74**(5): 399-407.

Lee, S., S.-J. Song, S. Koh, S. K. Lee and H.-g. Kim (2010). National Medical Terminology Server in Korea. Security-Enriched Urban Computing and Smart Grid, Springer**:** 541-544.

Lezcano, L., M.-A. Sicilia and C. Rodríguez-Solano (2011). "Integrating reasoning and clinical archetypes using OWL ontologies and SWRL rules." Journal of biomedical informatics **44**(2): 343-353.

Lin, J., D. Karakos, D. Demner-Fushman and S. Khudanpur (2006). Generative content models for structural analysis of medical abstracts. BioNLP Workshop on Linking Natural Language Processing and Biology at HLT-NAACL 06. New York City, USA, Association for Computational Linguistics**:** 65-72.

Literature, A. H. W. G. f. C. A. o. t. M. (1987). "A proposal for more informative abstracts of clinical articles." Annals of internal medicine **106**(4): 598-604.

Lorenzo Salazar, D. J. (2011). Lexical bundles in scientific English: A corpus-based study of native and non-native writing, Universitat de Barcelona.

Martínez-Costa, C., M. Menárguez-Tortosa, J. T. Fernández-Breis and J. A. Maldonado (2009). "A model-driven approach for representing clinical archetypes for Semantic Web environments." Journal of biomedical informatics **42**(1): 150-164.

McKnight, L. and P. Srinivasan (2003). "Categorization of sentence types in medical abstracts." AMIA Annu Symp Proc: 440-444.

Nakayama, T., N. Hirai, S. Yamazaki and M. Naito (2005). "Adoption of structured abstracts by general medical journals and format for a structured abstract." Journal of the Medical Library Association **93**(2): 237.

NLM. (2012). "Structured Abstracts in MEDLINE: Implementation Information." Retrieved September, 22, 2012, from http://structuredabstracts.nlm.nih.gov/Implementation.shtml.

Pathak, J., H. R. Solbrig, J. D. Buntrock, T. M. Johnson and C. G. Chute (2009). "LexGrid: a framework for representing, storing, and

querying biomedical terminologies from simple to sublime." Journal of the American Medical Informatics Association **16**(3): 305-315.

Perry, M., P. Jain and A. P. Sheth (2011). Sparql-st: Extending sparql to support spatiotemporal queries. Geospatial semantics and the semantic web, Springer**:** 61-86.

Pesquita, C., D. Faria, A. O. Falcao, P. Lord and F. M. Couto (2009). "Semantic similarity in biomedical ontologies." PLoS computational biology **5**(7): e1000443.

Ripple, A. M., J. G. Mork, L. S. Knecht and B. L. Humphreys (2011). "A retrospective cohort study of structured abstracts in MEDLINE, 1992-2006." Journal of the Medical Library Association **99**(2): 160-163.

Ripple, A. M., J. G. Mork, J. M. Rozier and L. S. Knecht (2012) "Structured Abstracts in MEDLINE: Twenty-Five Years Later."

Ron Daniel, J. (2012). "Domain-Independent Mining of Abstracts Using Indicator Phrases." D-Lib Magazine **18**(7/8).

Ruch, P., C. Boyer, C. Chichester, I. Tbahriti, A. Geissbühler, P. Fabry, J. Gobeill, V. Pillet, D. Rebholz-Schuhmann, C. Lovis and A.-L. Veuthey (2007). "Using argumentation to extract key sentences from biomedical abstracts." International Journal of Medical Informatics **76**(2–3): 195-200.

Smith, L., T. Rindflesch and W. J. Wilbur (2004). "MedPost: a part-of-speech tagger for bioMedical text." Bioinformatics **20**(14): 2320-2321.

Sollaci, L. and M. Pereira (2004). "The introduction, methods, results, and discussion (IMRAD) structure: a fifty-year survey." J Med Libr Assoc **92**: 364 - 367.

Sun, W., A. Rumshisky and O. Uzuner (2013). "Temporal reasoning over clinical text: the state of the art." Journal of the American Medical Informatics Association: amiajnl-2013-001760.

Swanson, D. R. (2001). "ASIST Award of Merit Acceptance Speech: On the Fragmentation of Knowledge, the Connection Explosion, and Assembling Other People's Ideas." Bulletin of the American Society for Information Science and Technology **27**(3): 12-14.

Tao, C., C. G. Parker, T. A. Oniki, J. Pathak, S. M. Huff and C. G. Chute (2011). An OWL meta-ontology for representing the clinical element model. AMIA annual symposium proceedings, American Medical Informatics Association.

van_der_Tol, M. (2001). "Abstracts as orientation tools in a modular electronic environment." Document Design **2**(1): 76-88.

Vawdrey, D. K. (2008). Assessing usage patterns of electronic clinical documentation templates. AMIA annual symposium proceedings, American Medical Informatics Association.

w3c (11. Feb. 2004). "Web Services Architecture."

W3C (2007. 27. April). "SOAP Version 1.2 Part 1: Messaging Framework (Second Edition)."

Ware, M. and M. Mabe (2015). "The STM report." An overview of scientific and scholarly journal publishing.

Wilbur, W. J., A. Rzhetsky and H. Shatkay (2006). "New directions in biomedical text annotation: definitions, guidelines and corpus construction." BMC bioinformatics **7**(1): 356.

Williams, I. A. (1996). "A contextual study of lexical verbs in two types of medical research report: Clinical and experimental." English for Specific Purposes **15**(3): 175-197.

Xu, R., K. Supekar, Y. Huang, A. Das and A. Garber (2006). "Combining text classification and hidden Markov modeling techniques for structuring randomized clinical trial abstracts." AMIA Annual Symposium Proceedings **2006**: 824-828.

Yamamoto, Y. and T. Takagi (2005). A Sentence Classification System for Multi Biomedical Literature Summarization. Data Engineering Workshops, 2005. 21st International Conference on.

Zhang, C. and X. Liu (2011). "Review of James Hartley's research on structured abstracts." Journal of Information Science **37**(6): 570-576.

Zweigenbaum, P., D. Demner-Fushman, H. Yu and K. B. Cohen (2007). "Frontiers of biomedical text mining: current progress." Briefings in bioinformatics **8**(5): 358-375.

사공철, 김종천 and 한국도서관협회 (1996). 문헌정보학용어사전, 한국도서관협회.

이경진 (2010). "[기획특집: E-잉크] 의료분야의 전자의무기록과 기록용 단말기 응용." Korean Industrial Chemistry News **13**(3): 14-22.

부록

Appendix 1 CDT 온톨로지

```
<?xml version="1.0" encoding="UTF-8"?>
<rdf:RDF xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
 xmlns:owl="http://www.w3.org/2002/07/owl#"
 xmlns:dcterms="http://purl.org/dc/terms/"
 xmlns:vann="http://purl.org/vocab/vann/"
 xmlns:foaf="http://xmlns.com/foaf/0.1/"
 xmlns:dc="http://purl.org/dc/elements/1.1/"
 xmlns:rdfs="http://www.w3.org/2000/01/rdf-schema#">

 <owl:Ontology rdf:about="http://vocab.bike.re.kr/cdt">
  <dcterms:title>CDT Ontology</dcterms:title>
  <dcterms:description>  CDT (Clinical Document Template) Ontology is an ontology for
  describing both structural and semantics-based clinical knowledge embedded
  in the level of clinical document templates. You can always find the latest
  version of the ontology at: https://github.com/SNUBiKE/CDT-Ontology
  </dcterms:description>
  <dcterms:license rdf:resource="http://creativecommons.org/licenses/by/3.0/"/>
  <dcterms:modified rdf:datatype="http://www.w3.org/2001/XMLSchema#date">2013-05-13</dcterms:modified>
  <vann:preferredNamespaceUri>http://vocab.bike.re.kr/cdt#</vann:preferredNamespaceUri>
  <vann:preferredNamespacePrefix>cdt</vann:preferredNamespacePrefix>
  <foaf:homepage rdf:resource="http://vocab.bike.re.kr/cdt.html"/>
  <dcterms:created rdf:datatype="http://www.w3.org/2001/XMLSchema#date">2012-07-22</dcterms:created>
  <dcterms:publisher rdf:resource="http://vocab.bike.re.kr/cdt#BiKE%2C%20SNU"/>
  <dcterms:partOf rdf:resource="http://vocab.bike.re.kr"/>
  <dcterms:type rdf:resource="http://purl.org/adms/assettype/Ontology"/>
  <dcterms:status rdf:resource="http://purl.org/adms/status/UnderDevelopment"/>
  <dc:creator rdf:resource="http://vocab.bike.re.kr/cdt#jgkim"/>
  <dc:creator rdf:resource="http://vocab.bike.re.kr/cdt#sjnam"/>
 </owl:Ontology>

 <rdf:Description rdf:about="http://vocab.bike.re.kr/cdt#ttl">
  <dcterms:license rdf:resource="http://creativecommons.org/licenses/by/3.0/"/>
 </rdf:Description>

 <rdf:Description rdf:about="http://vocab.bike.re.kr/cdt#rdf">
  <dcterms:license rdf:resource="http://creativecommons.org/licenses/by/3.0/"/>
 </rdf:Description>

 <foaf:Person rdf:about="http://vocab.bike.re.kr/cdt#jgkim">
  <foaf:name>James G. Kim</foaf:name>
  <foaf:homepage rdf:resource="http://jayg.org/"/>
 </foaf:Person>

 <dcterms:Agent rdf:about="http://vocab.bike.re.kr/cdt#BiKE%2C%20SNU">
  <foaf:member rdf:resource="http://vocab.bike.re.kr/cdt#jgkim"/>
  <foaf:member rdf:resource="http://vocab.bike.re.kr/cdt#sjnam"/>
  <foaf:name>BiKE, SNU</foaf:name>
  <foaf:homepage rdf:resource="http://bike.re.kr/"/>
 </dcterms:Agent>

 <foaf:Person rdf:about="http://vocab.bike.re.kr/cdt#sjnam">
  <foaf:name>Sejin Nam</foaf:name>
 </foaf:Person>

 <rdfs:Class rdf:about="http://vocab.bike.re.kr/cdt#AdjacentToRelation">
  <rdf:type rdf:resource="http://www.w3.org/2002/07/owl#Class"/>
  <rdfs:label>AdjacentTo Relation</rdfs:label>
  <rdfs:comment>
  A AdjacentTo Relation is a TDE relation that represents dispositional
  nextness of template description entities.
  </rdfs:comment>
  <rdfs:subClassOf rdf:resource="http://vocab.bike.re.kr/cdt#TDERelation"/>
 </rdfs:Class>

 <rdfs:Class rdf:about="http://vocab.bike.re.kr/cdt#CDTRelation">
  <rdf:type rdf:resource="http://www.w3.org/2002/07/owl#Class"/>
  <rdfs:label>CDT Relation</rdfs:label>
```

```
  <rdfs:comment>
  A CDT Relation is a relation from template components to clinical document
  templates.
  </rdfs:comment>
  <rdfs:subClassOf rdf:resource="http://vocab.bike.re.kr/cdt#Relation"/>
  <rdfs:subClassOf rdf:nodeID="arcfe34b1"/>
  <rdfs:subClassOf rdf:nodeID="arcfe34b2"/>
</rdfs:Class>

<rdfs:Class rdf:about="http://vocab.bike.re.kr/cdt#ClinicalDocumentTemplate">
  <rdf:type rdf:resource="http://www.w3.org/2002/07/owl#Class"/>
  <rdfs:label>Clinical Document Template</rdfs:label>
  <rdfs:comment>
  A Clinical Document Template (CDT) is a container for the content of a
  clinical document.
  </rdfs:comment>
</rdfs:Class>

<rdfs:Class rdf:about="http://vocab.bike.re.kr/cdt#Constraint">
  <rdf:type rdf:resource="http://www.w3.org/2002/07/owl#Class"/>
  <rdfs:label>Constraint</rdfs:label>
  <rdfs:comment>
  A Constraint is a concept that defines a range of values a template
  description entity can posesses, or specifies generic constraints among
  template description entities.
  </rdfs:comment>
</rdfs:Class>

<rdfs:Class rdf:about="http://vocab.bike.re.kr/cdt#ContainedInRelation">
  <rdf:type rdf:resource="http://www.w3.org/2002/07/owl#Class"/>
  <rdfs:label>ContainedIn Relation</rdfs:label>
  <rdfs:comment>    A ContainedIn Relation is a TDE relation that represents template
  description entities' belonging to other template description entities.
  </rdfs:comment>
  <rdfs:subClassOf rdf:resource="http://vocab.bike.re.kr/cdt#TDERelation"/>
</rdfs:Class>

<rdfs:Class rdf:about="http://vocab.bike.re.kr/cdt#Context">
  <rdf:type rdf:resource="http://www.w3.org/2002/07/owl#Class"/>
  <rdfs:label>Context</rdfs:label>
  <rdfs:comment>
  A Context is a concept that represents circumstances in which a template
  description entity appears.
  </rdfs:comment>
</rdfs:Class>

<rdfs:Class rdf:about="http://vocab.bike.re.kr/cdt#Relation">
  <rdf:type rdf:resource="http://www.w3.org/2002/07/owl#Class"/>
  <rdfs:label>Relation</rdfs:label>
  <rdfs:comment>
  Relation is a meta concept that subsumes the different types of binary or
  n-ary relations that can be defined.
  </rdfs:comment>
</rdfs:Class>

<rdfs:Class rdf:about="http://vocab.bike.re.kr/cdt#TDERelation">
  <rdf:type rdf:resource="http://www.w3.org/2002/07/owl#Class"/>
  <rdfs:label>TDE Relation</rdfs:label>
  <rdfs:comment>
  A TDE Relation is a relation among template description entities.
  </rdfs:comment>
  <rdfs:subClassOf rdf:resource="http://vocab.bike.re.kr/cdt#Relation"/>
  <rdfs:subClassOf rdf:resource="http://vocab.bike.re.kr/cdt#TemplateComponent"/>
  <rdfs:subClassOf rdf:nodeID="arcfe34b3"/>
  <rdfs:subClassOf rdf:nodeID="arcfe34b4"/>
</rdfs:Class>

<rdfs:Class rdf:about="http://vocab.bike.re.kr/cdt#TemplateComponent">
  <rdf:type rdf:resource="http://www.w3.org/2002/07/owl#Class"/>
  <rdfs:label>Template Component</rdfs:label>
  <rdfs:comment>
  A Template Component is something that can be in a clinical document
  template.
  </rdfs:comment>
</rdfs:Class>

<rdfs:Class rdf:about="http://vocab.bike.re.kr/cdt#TemplateDescriptionEntity">
  <rdf:type rdf:resource="http://www.w3.org/2002/07/owl#Class"/>
```

```
  <rdfs:label>Template Description Entity</rdfs:label>
  <rdfs:comment>
  A Template Description Entity (TDE) is a data entity holding a key/value
  pair in a clinical document template.
  </rdfs:comment>
  <rdfs:subClassOf rdf:resource="http://vocab.bike.re.kr/cdt#TemplateComponent"/>
</rdfs:Class>

<rdfs:Class rdf:about="http://vocab.bike.re.kr/cdt#UsedAtRelation">
  <rdf:type rdf:resource="http://www.w3.org/2002/07/owl#Class"/>
  <rdfs:label>UsedAt Relation</rdfs:label>
  <rdfs:comment>
  A UsedAt Relation is a CDT relation that represents usage of template
  components in clinical document templates.
  </rdfs:comment>
  <rdfs:subClassOf rdf:resource="http://vocab.bike.re.kr/cdt#CDTRelation"/>
</rdfs:Class>

<rdf:Property rdf:about="http://vocab.bike.re.kr/cdt#department">
  <rdf:type rdf:resource="http://www.w3.org/2002/07/owl#DatatypeProperty"/>
  <rdfs:label>department</rdfs:label>
  <rdfs:comment>
  The department where the clinical document template is defined and used.
  </rdfs:comment>
  <rdfs:domain rdf:resource="http://vocab.bike.re.kr/cdt#ClinicalDocumentTemplate"/>
  <rdfs:range rdf:resource="http://www.w3.org/2000/01/rdf-schema#Literal"/>
</rdf:Property>

<rdf:Property rdf:about="http://vocab.bike.re.kr/cdt#filename">
  <rdf:type rdf:resource="http://www.w3.org/2002/07/owl#DatatypeProperty"/>
  <rdfs:label>filename</rdfs:label>
  <rdfs:comment>
  The name of the file in which the clinical document template is stored.
  </rdfs:comment>
  <rdfs:domain rdf:resource="http://vocab.bike.re.kr/cdt#ClinicalDocumentTemplate"/>
  <rdfs:range rdf:resource="http://www.w3.org/2000/01/rdf-schema#Literal"/>
</rdf:Property>

<rdf:Property rdf:about="http://vocab.bike.re.kr/cdt#constraint">
  <rdfs:label>constraint</rdfs:label>
  <rdfs:comment>
  Links to a constraint the template description entity has.
  </rdfs:comment>
  <rdfs:domain rdf:resource="http://vocab.bike.re.kr/cdt#TemplateDescriptionEntity"/>
  <rdfs:range rdf:resource="http://vocab.bike.re.kr/cdt#Constraint"/>
</rdf:Property>

<rdf:Property rdf:about="http://vocab.bike.re.kr/cdt#context">
  <rdfs:label>context</rdfs:label>
  <rdfs:comment>
  Links to a context in which the template description entity appears.
  </rdfs:comment>
  <rdfs:domain rdf:resource="http://vocab.bike.re.kr/cdt#TemplateDescriptionEntity"/>
  <rdfs:range rdf:resource="http://vocab.bike.re.kr/cdt#Context"/>
</rdf:Property>

<rdf:Property rdf:about="http://vocab.bike.re.kr/cdt#source">
  <rdfs:label>source</rdfs:label>
  <rdfs:comment>
  Links to an instance that is a source or origin of the directed binary
  or n-ary relation.
  </rdfs:comment>
  <rdfs:domain rdf:resource="http://vocab.bike.re.kr/cdt#Relation"/>
</rdf:Property>

<rdf:Property rdf:about="http://vocab.bike.re.kr/cdt#target">
  <rdfs:label>target</rdfs:label>
  <rdfs:comment>
  Links to an instance that is a target of the directed binary or n-ary
  relation.
  </rdfs:comment>
  <rdfs:domain rdf:resource="http://vocab.bike.re.kr/cdt#Relation"/>
</rdf:Property>

<owl:Restriction rdf:nodeID="arcfe34b1">
  <owl:onProperty rdf:resource="http://vocab.bike.re.kr/cdt#source"/>
  <owl:allValuesFrom rdf:resource="http://vocab.bike.re.kr/cdt#TemplateComponent"/>
</owl:Restriction>
```

```
  <owl:Restriction rdf:nodeID="arcfe34b2">
    <owl:onProperty rdf:resource="http://vocab.bike.re.kr/cdt#target"/>
    <owl:allValuesFrom rdf:resource="http://vocab.bike.re.kr/cdt#ClinicalDocumentTemplate"/>
  </owl:Restriction>

  <owl:Restriction rdf:nodeID="arcfe34b3">
    <owl:onProperty rdf:resource="http://vocab.bike.re.kr/cdt#source"/>
    <owl:allValuesFrom rdf:resource="http://vocab.bike.re.kr/cdt#TemplateDescriptionEntity"/>
  </owl:Restriction>

  <owl:Restriction rdf:nodeID="arcfe34b4">
    <owl:onProperty rdf:resource="http://vocab.bike.re.kr/cdt#target"/>
    <owl:allValuesFrom rdf:resource="http://vocab.bike.re.kr/cdt#TemplateDescriptionEntity"/>
  </owl:Restriction>

  <rdf:Description rdf:about="http://purl.org/dc/terms/created">
    <rdfs:subPropertyOf rdf:resource="http://www.w3.org/ns/prov#generatedAtTime"/>
  </rdf:Description>

  <rdf:Description rdf:about="http://purl.org/dc/terms/creator">
    <rdfs:subPropertyOf rdf:resource="http://www.w3.org/ns/prov#wasAttributedTo"/>
  </rdf:Description>

  <rdf:Description rdf:about="http://purl.org/dc/terms/modified">
    <rdfs:subPropertyOf rdf:resource="http://www.w3.org/ns/prov#generatedAtTime"/>
  </rdf:Description>

  <rdf:Description rdf:about="http://purl.org/dc/terms/references">
    <rdfs:subPropertyOf rdf:resource="http://www.w3.org/ns/prov#wasDerivedFrom"/>
  </rdf:Description>

</rdf:RDF>
```

Abstract

The Study on Automatic Annotation using Structural/Linguistic Characteristics of biomedical documents

Se-Jin Nam
Healthcare Management and Informatics
The Graduate School
Seoul National University

There has been a rapid increase in research of automatic annotation for biomedical articles and clinical documents. With this increaseit is important to provide the ability to search and extract information from these documents for biomedical researchers and clinicians. This research starts with the needed annotation techniques that make it possible for biomedical researchers to search for documents, and clinicians to search for informationto make a diagnosis and write prescription records. These two activities, searching, and recording for biomedical articles and clinical document templates is a common occurrence in the biomedical domain. This is why it is important to improve the effectiveness for these two activities.

An abstract plays an important role in summarizing what an article is

about. That is why this research first started with automatic tagging of unstructured abstracts using the popular IMRAD structure in biomedicine. Methods from both linguistics and computer science were used on biomedical articles. Using these, we developed an automatic tagging system that had a high performance with a low computational cost. For clinical documents, since most of them are created from clinical document templates to be used in EMR systems, we started with the automatic tagging of clinical document templates. Clinical document templates are different from most research articles because they are already structured. This study focuses on the tagging of the inherent knowledge in these documents. We propose a new knowledge model with STEP, a system that uses this model to accomplish our goal.

Using the proposed methods, we developed an automatic tagging system that had a high performance with a low computational cost. With only 17 features extracted from documents, the system was able to classify unstructured abstracts with an accuracy of 77.0~ 90.3%. With additional features, the system had an accuracy of up to 91.7%. In case of STEP, to test the effectiveness of STEP, a clinical document template system was developed. We show how a constructed knowledge base with our proposed knowledge model can aid intelligently with filling in clinical document templates.

This study, focused on the two representative types of documents in biomedicine, is significant in that it provides a technique to improve search and management of two documents efficiently and a base technique that could be used in text mining and applied systems for biomedical domain. We also provide a web application for other

researchers to utilize the results of our research.

Keywords: Annotation, Structured abstracts, Clinical documents, Setence classification, Ontology

Student Number: 2010-30648